Global dissection of the BAHD acyltransferase gene family in soybean: Expression profiling, metabolic functions, and evolution

Muhammad Zulfiqar Ahmad
Anhui Agricultural University https://orcid.org/0000-0001-7201-8937
Xiangsheng Zeng
Anhui Agricultural University
Qiang Dong
Huazhong Agriculture University
Sehrish Manan
Huazhong Agriculture University
Huanan Jin
Huazhong Agriculture University
Penghui Li
Anhui Agricultural University
Xiaobo Wang
Anhui Agricultural University
Vagner A. Benedito
West Virginia State University
Jian Zhao (jianzhao@ahau.edu.cn)
https://orcid.org/0000-0002-4416-7334

Research article

Keywords: BAHD, Domestication, Evolution, Gene expression, Genome, Positive Selection, Soybean, Stress

Posted Date: January 22nd, 2020

DOI: https://doi.org/10.21203/rs.2.21482/v1

Abstract

Background: Members of the BAHD acyltransferase (ACT) family play important roles in plant defence against biotic and abiotic stresses. Previous genome-wide studies explored different acyltransferase gene families, but not a single study was found so far on the overall genome-wide or positive selection analyses of the BAHD family genes in Glycine max . A better understanding of the functions that specific members of this family play in stress defence can lead to better breeding strategies for stress tolerance.

Results: A total of 103 genes of the BAHD family (GmACT genes) were mined from the soybean genome, which could be grouped into four phylogenetic clades (I- IV). Clade III was further divided into two sub-clades (IIIA and IIIB). In each clade, the constituent part of the gene structures and motifs were relatively conserved. These 103 genes were distributed unequally on all 20 chromosomes, and 16 paralogous pairs were found within the family. Positive selection analysis revealed important amino acids under strong positive selection, which suggests that the evolution of this gene family modulated soybean domestication. Most of the expression of ACT genes in soybean was repressed with Al 3+ and fungal elicitor exposure, except for GmACT84 , which expression increased in these conditions 2- and 3-fold, respectively. The promoter region of GmACT84 contains the maximum number of stress-responsive elements among all GmACT genes and is especially enriched in MYB-related elements. Some GmACT genes showed expression specific under specific conditions, while others showed constitutive expression in all soybean tissues or conditions analysed.

Conclusions: This study provided a genome-wide analysis of the BAHD gene family and assessed their expression profiles. We found evidence of a strong positive selection of GmACT genes. Our findings will help efforts of functional characterisation of ACT genes in soybean in order to discover their involvement in growth, development, and defence mechanisms.

Background

A large number of secondary metabolites in plants play significant roles in their response to environmental challenges [65]. The specialised metabolites present in a diverse range of the plant kingdom are synthesised through the alteration of the basic structure by different reactions, such as acylation [9], methylation [39], and glycosylation [10]. E.g., these molecules are involved in seed and flower colouration to help pollination and dispersion [14]. Modifications of phenolic compounds through acylation enhances their functions in planta. For example, acylation to flavones, isoflavone or anthocyanins has a substantial impact on their solubilization, stabilization, transport, vacuolar uptake, and colour modulation [56, 88–89]. The sinapoyl moiety conjugation to malate enhances the anti-UV property of flavonoids [40]. Avenacin acylation plays a role in plant defence [52]. The acylation of anthocyanin, flavones or isoflavones increases their solubility, transport and storage properties [26, 90].

The BAHD family comprises of a large acyltransferase (ACT) gene family found in plants and prokaryotes. They use CoA conjugates (e.g., acetyl-, malonyl-, β-phenylalanine, tiglyl, anthraniloyl, and

benzoyl- moieties) as donors to modify different acceptor compounds [9, 16]. The O- or N-acylation reaction is carried out by BAHD acyltransferases via the acceptor's hydroxyl (OH) or amine (NH_2) group with different CoA donor substrates. The synthesis in plants of a variety of polymers and secondary metabolites like lignin, volatiles, suberin, cutin, pigments, and defence-related compounds involves these enzymes [16]. Furthermore, several BAHD genes that catalyse the formation of diverse metabolites have been characterised in different plant species. For example, E. coli was used to express several BAHD enzymes, such as GmIMaT1, GmIMaT3 [2], GmMT7 [20], MtMaT1, 2, 3 [86], MaT4, 5, 6 [89] for the synthesis of malonylated flavones, isoflavones or anthocyanins, HQT for the synthesis of the antioxidant chlorogenic acid [33], hydroxycinnamoyl-CoA:hydroxyphenyllactate transferases for the production of rosmarinic acid [8], the hydroxycinnamoyl-CoA:glycerol transferase for the production of water-soluble antioxidants, hydroxycinnamate glycerol esters [34], and the hydroxycinnamoyl/benzoyl-CoA:anthranilate transferase (HCBT) for the synthesis of therapeutic benzoyl and hydroxycinnamoyl anthranilates [22]. The completion of genome sequencing projects of many plant species revealed the presence of BAHD family genes, such as the models Arabidopsis thaliana and Oryza sativa as well as orphan species, such as the cardoon (Cynara cardunculus, Asteraceae). Their genomes contain 64, 119, and 69 BAHD genes, respectively [16–17, 49]. Up to now, there have been no genome-wide analyses on the BAHD family in the model legume crop, Glycine max.

Although G. max is an important crop worldwide, little work on BAHD family genes has been carried out on this species when compared to other model and crop plant species. To date, only two BAHD genes were reported in soybean: GmIMaT1 and the two alleles of GmIMaT3 [2], GmMT7 [20] and GmIF7MaT [75], which respectively malonylate the isoflavones daidzin, genistin, and glycitin. The whole-genome sequence of soybean is publicly available at Phytozome and soyKB databases [67, 70], making it possible to identify and analyse the BAHD genes as well as explore their expression patterns and create strong hypotheses on their metabolic functions. Sequence comparisons of BAHD genes in soybean with already characterised genes in other species will assist gene function predictions. In this study, we have executed a global analysis of all BAHD genes in the soybean genome. The chromosomal distribution, phylogeny, gene structure, duplication events, positive selection, and transcriptional dynamics in response to different stresses were analysed in this gene family.

Results And Discussion

Identification of BAHD family members in the soybean genome

Considering only the genes coding the specific BAHD catalytic motif (HXXXD), we identified 103 genes (Additional files 1: File S3) in the soybean genome encoding proteins of the BAHD family. The annotation of soybean BAHD genes was based on their fall in the clade and position on the chromosome (GmACT1–99) along with the already characterised genes Glyma.18G268200 (GmIMaT1), Glyma.13G056100 (GmIMaT3) [2], Glyma.18G258000 (GmMaT2) and Glyma.18G268400 (GmMaT4) (Fig. 1A). Details about

on physical parameters, such as Glyma gene ID, chromosome start-end position, coding strand, exon number, predicted protein sequence length, molecular weight, isoelectric point (pI), and predicted subcellular location are listed in Additional files 1: File S3. Predicted protein length varied from 274 (Glyma.13G007200_GmACT31) to 511 (Glyma.18G258000_GmMaT2) amino acid residues, which indicates high variation within members of the BAHD family. However, most of them (70 out of the 103 members) code a protein with 450–520 residues. The range of predicted molecular weight is from 30.93–56.70 kD whereas the majority of the genes code a product with a predicted molecular weight ranging 45–52 kD, which falls within the values of characterised BAHD genes [16]. The predicted pI value of soybean BAHD proteins ranged from acidic 5.05 (Glyma.10G061900_GmACT53) to basic 9.27 (Glyma.02G264600_GmACT2) (Additional files 1: File S3). The predicted subcellular localization indicates that BAHD proteins could be present in almost all organelles of the soybean cell. However, GmIMaT1 and GmIMaT3 are primarily localised in endoplasmic reticulum and cytosol, respectively [2]. Most BAHD proteins (37/103) were predicted to localize in the cytosol, followed by mitochondria (27/103) and the plasma membrane (20/103), whereas only 2 proteins are predicted to localize in the nucleus (Additional files 1: File S3).

Motif analysis and gene structure

Variation in length and sequence was found in the conserved region of soybean BAHD genes (Additional files 1: File S4). Among the ten representative BAHD motifs searched and annotated through Pfam and SMART tools, three motifs were always present in each gene identified (Additional files 1: File S4). The biological importance of these recurring motifs required an in-depth investigation. All clade I and clade II members contained all 10 conserved motifs while clade III and IV members missed some of the motifs (Fig. 1B). We found a typical motif pattern in most of the paralogous pairs. However, some pairs presented differences in motif composition and arrangement, such as GmACT94-GmACT99, GmACT93-GmACT96, GmACT31-GmACT46, GmACT21-GmACT22, GmACT52-GmACT72, GmACT7-GmACT9 and GmACT15-GmACT18 (Fig. 1B).

The gene architecture of soybean BAHD genes was analyzed to better understand the general features within each clade. The exon-intron structure of each gene is shown along with the phylogenetic tree (Fig. 1C). Genes within the same clade contain almost same length and number of exons and introns, with some exceptions (Fig. 1C). Introns of most clade I members are quite larger than of those members in other clades. Seventy percent (14 out of 20) of clade I members contain two exons while 15% (3/20) contain one and three exons (Fig. 1C; Additional files 1: File S3). GmACT12 contains the longest intron in the whole set of soybean BAHD genes (Fig. 1C). In clade II, 14 genes contain only one, 8 genes contain two, and 3 genes contain three exons. GmACT21, GmACT22, GmACT32 and GmACT33 have long introns compared to the other clade II members. Clade III genes varied in exon number, but most contain just a single exon. Two clade III members (GmACT51 and GmACT72) contain the maximum exon number (4) in whole soybean BAHD family. Most members of clade IV (76%) contain only one exon while the others contain two exons. Among the 16 paralogous pairs, 10 did not show remarkable differences except for GmACT21-GmACT22 of clade II which a member of the pair has an untranslated region (UTR) longer than

the other member. The other six paralogous pairs showed small variations in exon number or intron length (Fig. 1C).

Phylogenetic relationships of the soybean BAHD members

The evolutionary relationship of the soybean BAHD genes was explored by constructing a phylogenetic tree to compare them with 43 functionally characterised BAHD genes from different plant species (Fig. 2). On the basis of substrate donor and acceptor, the 103 soybean BAHD genes were classified into four major clades (I-IV) and clade III was further subdivided into two subclades (IIIA and IIIB) (Fig. 2). Our results based on the phylogenetic relationships of the characterised soybean ACT proteins would predict their correct enzymatic functions.

Clade I contains 20 soybean BAHD genes along with 16 characterised genes from different plant species (Fig. 1). Most members of this clade code for benzoyltransferases, alkyl-hydroxycinnamate ester, or cell-wall acyltransferases. They use benzoyl as a donor substrate to form benzyl benzoates, such as CbBEBT, NtBEBT [18], or salicyl benzoate (PtSABT) and involved in plant defence mechanisms against herbivory [12]. Many characterised BAHD enzymes of clade I are responsible for the formation of benzenoid compounds [7, 12, 18]. Among them, PtSABT from Populus trichocarpa uses different CoA thioesters and alcohols as a donor to catalyse the production of salicyl benzoate for the synthesis of a variety of acetates, benzoates and cinnamates [12]. PtBEBT from Populus trichocarpa is more selective to produce benzyl benzoate [12] and takes part in the synthesis of salicinoids. CbBEBT produces benzyl benzoate in Clarkia breweri that composes its floral aroma. The CbBEBT gene is highly expressed in the flower, and especially in the stigma [18]. These BEBT genes code for products that have the capability to produce compounds that are important for plant defence or aroma [7, 12, 18]. Four BEBT genes from soybean (GmACT3, GmACT4, GmACT5, GmACT20) have high similarity with the characterised BEBT genes. Similarly to CbBEBT, GmACT3 and GmACT4 also showed high expression levels in the flower, which leads us to hypothesize that it likely plays a function in catalyzing aroma compounds. These genes are also highly expressed in the stem and the leaf, like PtSABT and PtBEBT that produce defence compounds. Some members of clade I were also reported being involved in the biosynthesis of cutin and suberin, which are known to be lipid barriers on organ epidermis surfaces. These barriers regulate gas, water and solute fluxes [63] and play critical roles against pathogen attacks [9]. Alkaylhydroxycinnamoyl ester BAHD genes, such as AtASFT [50], StFHT [69], AtDCF [64], and AtFACT [37] use feruloyl, p-coumaroyl, sinapoyl, and caffeoyl as a donor to form aromatic suberin, alkyl, or ω-hydroxy fatty acids, which are used for the synthesis of cutin polymers that in turn are important components of the cuticle and assist in the deposition of suberin into the inner side of the primary cell wall of secondary shoots and roots [9]. They also play a role in the synthesis of root waxes that contain hydroxycinnamates in A. thaliana [37]. Ten genes from soybean fell in clade I and, thus have the same predicted transferase function (Fig. 2; Additional files 1: File S3). Interestingly, we have found in soybean only one orthologue of disinapoylspermidine (AtSDT; [44]) and another of coumaroyl-spermidine (AtSCT; [44]) that were annotated as GmACT2 and GmACT12, respectively (Fig. 2) while all the other genes in this clade have multiple copies or paralogous in the soybean genome. Monolignols, which participate in the formation of

lignin, are synthesised through the phenylpropanoid pathway, primarily through the activity of hydroxycinnamoyltransferases (HCTs). The BAHD-ACTs also used the hydroxycinnamic acid (HCA) as donor substrate to esterify monolignols [84]. These cell wall-related BAHD-ACTs also belong to clade I, such as OsPMT [84], BdPMT [62], and OsAT10 [5] that use p-coumaroyl-CoA as a donor substrate to form lignols and glucuronoarabinoxylan (GAX). These activities are responsible for producing the lignocellulosic biomass from grasses, which could be a major feedstock for the generation of liquid biofuels [62]. The soybean genes GmACT1, GmACT13, GmACT14, GmACT16, GmACT17, and GmACT19 were also predicted to have cell wall-related functions, and they may play essential roles in the production of liquid biofuel compounds.

Clade II of the soybean BAHD family contains 26 genes and cluster together with 9 characterised from other species (Fig. 2). This clade includes many genes related to shikimate/quinate hydroxycinnamoyltransferases (HCT/HQT) and acyl-CoA N-acyltransferases, such as hydroxycinnamoyl-CoA, spermidinehydroxycinnamoyltransferase (SHT), hydroxycinnamoyl-CoA: N-hydroxycinnamoyl/benzoyltransferase (HCBT), and rosmarinic acid synthase, which is alternatively called hydroxycinnamoyl-CoA:hydroxyphenyllactate hydroxycinnamoyltransferase (RAS). These members used hydroxycinnamoyl, caffeoyl or p-coumaroyl as donor substrates to form chlorogenic acid, hydroxycinnamoylquinate, caffeoyl, or p-coumaroylquinate or malate, hydroxycinnamoyl glycerol, hydroxycinnamoylshikimate, hydroxyphenyllactates, anthramides, and spermidines [27, 29, 43, 58, 64, 74, 83]. Clade II contains mostly acyl-CoA N-ACTs that catalyze N-acylation instead of O-acylation. These reactions catalyze the hydroxycinnamoyl and amine moieties that are the entry point of the phenolamide pathway [9]. Hydroxycinnamic acid (HCA) amides are classified as phytoalexins [23]. Most of the plant N-HCTs that have been characterised so far are from A. thaliana and N. attenuate and involved in defence mechanism against herbivores [60]. While other members of the clade II, i.e., HCT/HQT and their relatives produce different compounds like lignin as well as chlorogenic acid (CGA) and quinates, which play protective roles against UV photooxidation [27] and pathogenic resistance [58]. CGA has also been implicated in cell wall development [51] and root hair formation [57]. Seventeen soybean BAHD genes are closely related to TpHCT2, which product transfers hydroxycinnamoyl moieties (p-coumaroyl and caffeoyl) to malic acid (Fig. 1). This transferase activity of TpHCT2 is involved in phaselic acid synthesis, particularly in the leaves of red clover (Trifolium pratense, Fabaceae) [74]. TpHCT2 is related to GmACT27, GmACT28, and GmACT45, which also have high and specific expression in leaves (Fig. 2), and may play a role in phaselic acid synthesis. Meanwhile, GmACT21, GmACT22, GmACT32, and GmACT33 fall closest with TpHCT1A/1B, which use caffeoyl or p-coumaroyl as a donor to form shikimates or malates [74] and play a major role in monolignol synthesis [29, 74]. Lignin precursors are synthesised through p-coumaroyl derivatives rather than methoxylated analogues in response to pathogen attacks [80]. TpHCT activity was higher in the stem and capable of transferring p-coumaroyl moieties from the respective CoA to shikimic acid, which is critical for the biosynthesis of monolignol lignin precursors [29, 74]. AtHCT expression dramatically changed the lignin composition, which demonstrates the essential function of HCT in phenylpropanoid metabolism [6, 29]. HCT homologues have been identified in several vascular plant species [9]. In conformity with other members of this group,

GmACT21, GmACT22, GmACT32 also presented high expression levels in the stem, and they may transfer p-coumaroyl moieties to respective acceptors with essential roles in lignin biosynthesis. Other four members of this clade (GmACT35, GmACT36, GmACT38, and GmACT39) are closely related to SHT from Arabidopsis, which uses hydroxycinnamoyl as a donor substrate to form hydroxycinnamoylspermidine [27]. This compound is abundant in pollen of many species, which indicates a role of these phenylpropanoids in the pollen [27]. AtSHT expression was high and specific to the tapetum cells of the anther during early flower development in order to supply the building blocks for cell wall development in the pollen [27]. The metabolic profiling of flower bud tissues in Arabidopsis revealed that AtSHT knockout or RNAi plants had a positive correlation between acyltransferase protein accumulation and the amounts of trihydroxyferuloyl and dihydroxyferuloyl sinapoyl spermidines [27]. GmACT35 and GmACT38 were also more highly expressed in the flower than any other tissues and are expected to have a function in the acylation of spermidine in the pollen.

Clade III is further divided into sub-clades IIIA and IIIB (Fig. 2). Sub-clade IIIA contains 25 genes along with 19 already characterised from other species. Most of the characterised members of clade IIIA acylate phenolic glucoside anthocyanins or other flavonoids (flavones or isoflavone). Malonylation of clade IIIA ACTs depends on the position of the acceptor substrate. Clade IIIA can be divided into aliphatic and aromatic ACTs. Fourteen soybean BAHD genes fall in the aliphatic ACTs, and they are functionally annotated as isoflavonoid malonyltransferases (Additional file 1: File S3) along with the characterised At5MaT, At3At1, At3At2 from Arabidopsis and MtMaT1, MtMaT2 and MtMaT4 from Medicago, which facilitate the transport or storage of the metabolites [44, 89]. Among the 14 aliphatic ACTs in soybean, only two, GmIMaT1 and GmIMaT3 [2] have been functionally characterised and play a distinct role against abiotic and hormonal stresses [2]. Their transcription patterns correlated positively with the malonylisoflavonoid composition of the tissue in different parts of the plant under different stress conditions [2].

Another aliphatic group contains 6 soybean ACTs along with MtMaT5 and MtMaT6 [89] from Medicago that malonylate different anthocyanins, such as cyanidin, delphinidin, and pelargonidin conjugates at 3-O-glucosides [89]. These genes are very specific for anthocyanin modification at the C3 position, and they have no activity towards C5 or C7 of anthocyanin or other flavones or isoflavones [89]. The soybean genes that closely relate to MtMaT5 and MtMaT6 are predicted to function in anthocyanin modification at 3-O-glucoside (Additional file 1: File S3). This group of clade IIIA is most probably associated with the stability of the anthocyanin glucosides since anthocyanin acylation at 3-O sugar with the malonyl moiety has been suggested to increase its stability [16].

Clade IIIA also contains 7 soybean BAHD-ACT genes (GmACT52, GmACT57, GmACT61, GmACT63, GmACT72, GmACT76 and GmACT77) closely associated with the aromatic ACTs from Arabidopsis, At3At1 and At3At2 (Additional file 1: File S3; [44]). At3At1 and At3At2 use p-coumaroyl, feruloyl and caffeoyl CoAs as donors to specifically modify the C6 hydroxyl of glucose at three anthocyanin positions, but it did not have any affinity toward sinapoyl-CoA. We predict that these 7 aromatic ACTs will also prefer coumaroyl-CoA to modify the anthocyanin at the C3 position just like their closely related

Arabidopsis orthologues (Fig. 2; Additional file 1: File S3). We did not find in the soybean genome any aromatic gene that was closely related to a gene able to modify 5-O-glucosides, such as Gt5AT that acylates the C6 hydroxyl of 5-O-glucosyl residues of anthocyanin with either caffeoyl, p-coumaroyl [25] or aliphatic anthocyanin MaTs, such as At5MaT (Fig. 2; [15]) and Ss5MaT1 from Arabidopsis thaliana (Brassicaceae) and Salvia splendens (Lamiaceae), respectively (Fig. 2; [77]).

Clade IIIB contains 9 GmACT genes from soybean with 7 characterised genes (Fig. 2). This clade contains malonyltransferases (MaTs) that catalyse flavonoids with malonyl-CoA to form malonates, acetyl CoA:deacetylvindoline 4-O-acetyltransferases (DAT) which catalyse the last step in vindoline biosynthesis [73], acylsugar acyltransferases (ASAT), which is important in the acylsucrose biosynthetic pathway and uses sucrose and acyl-CoA ester substrate specificity to contribute in different types of accumulated acylsucroses [24]. Acylsugars are polyesters of short- to medium-length acyl chains of sucrose or glucose backbones that are produced in secretary glandular trichomes of many solanaceous plants, including the cultivated tomato (Solanum lycopersicum). These compounds act against different insects through a mechanism involving in a multitrophic defence interaction involving their metabolization to volatile fatty acids [35]. Ss5MaT2 also fell in clade IIIB separated from other the MaTs, which is consistent with D'Auria analyses [16]. These BAHD-ACTs share the anthocyanin specific motif (NYFGNC) along with the BAHD HXXXD motif. The characterised member of clade IIIB (Ss5MaT2 from Salvia splendens) malonylates the anthocyanin glycosides bisdemalonylsalvianin and shisonin by transferring the malonyl moiety to the C4 position of the 5-O-glucosyl residue [76]. The soybean genes of clade IIIB also suggest forming salvianin, which has two malonyl moieties, and malonylshisonin, which plays an important role in flower colour [76]. Although the biochemical properties of Ss5MaT2 are similar to that of clade IIIA members, it fell into the clade with ASAT and DAT.

Clade IV contains 21 soybean BAHD genes along with five characterised genes (Fig. 2; Additional file 1: File S3). The members of this clade have varied functions. Most are predicted to function as shikimate O-hydroxycinnamoyltransferase or transferase family, except for GmACT88 and GmACT97, which are predicted to function as an anthranilate N-hydroxycinnamoyl (HCBT)-related transferase, as well as GmACT79 and GmACT81, which are predicted to function as a shikimate O-hydroxycinnamoyltransferase/quinate O-hydroxycinnamoyltransferases. Gt5At from Gentiana triflora, an anthocyanin aromatic acyltransferase, catalyses the transfer of hydroxycinnamic acid moieties from their CoA esters to the glycosyl groups of anthocyanins [25]. HCBT produces the aromatic amide dianthramide in carnation [83]. Anthraniloyl-CoA: methanol AT (AMAT) uses anthraniloyl-CoA as a donor substrate to acylate methanol and produce a foxy odour compound (methyl anthranilate) in concord grapes [81]. Moreover, GmACT93 and GmACT96, which were predicted to involved in cutin formation and closely related to the gene DEFECTIVE IN CUTICULAR RIDGES (DCR) from Arabidopsis thaliana, Medicago truncatula, and Populus trichocarpa [61]. In the flower, DCR converts monomers into polymeric cutin, which is composed of fatty acid, glycerol and aromatic monomers [32]. Changes in epidermal cell differentiation and postgenital organ fusion were observed due to defective cuticle in DCR-deficient plants. DCR was also expressed in the root cap, lateral root primordia, and developing lateral root. Its root expression might point to involvement in the biosynthesis of suberin [61]. Downregulation of DCR in dcr

mutants led to excessive root branching and changes of root architecture by influencing lateral root emergence and growth [61]. All in all, these analyses help us formulate educated hypotheses about the diversity and evolution of substrate specificity within the BAHD family in soybean.

Chromosomal localization and duplication of soybean BAHD genes

A chromosome ideogram was constructed on the basis of the physical position of the soybean BAHD-ACT genes (Additional file 1: File S3). All the 20 soybean chromosomes contain BAHD-ACT genes, although the number on each chromosome varies between 1 and 15 genes per chromosome (Fig. 3). Chromosome 18 has the highest number of genes (15%) followed by chromosomes 8 (11%) and 19 (10%) (Fig. 3; Additional file 1: File S3). On the other hand, chromosomes 1, 9 and 20 contain only one gene each (Fig. 3; Additional file 1: File S3). Given their position, often on chromosome arms (Fig. 3), soybean BAHD genes tend to have high rates of recombination [32].

Tandem and segmental duplications increased the gene numbers into a large family in different plant species [79]. Potential gene redundancies can be discovered by the identification of closely related paralogues in the genome whereas the accurate identification of true orthologues between species can lead to the creation of strong hypothesis of common gene function in other species. The soybean genome contains a high number (103) of BAHD genes as compared to 69 in Cynara cardunculus [49], 64 in A. thaliana [17], but less the 119 BAHD genes in the Oryza sativa genome [16]. This large gene family might be the result of the two rounds of whole-genome duplication (WGD) events that occurred 58–60 (the Papilionoideae allotetraploidization event) and 13 Mya (recent soybean allotetraploidization) in soybean [66]. The presence of paralogous genes in a plant genome is more often related to their functional category than the genetic proximity between species [67]. We found sixteen paralogous pairs in the soybean BAHD family (Additional file 2: Table S1). These paralogues are present in all the four phylogenetic clades (Additional file 2: Table S1). Clade I contains 7, clade II contains 3, clade III contains 2, and clade IV contains 4 paralogous pairs (Additional file 2: Table S1). The presence of clear paralogues is evidence of common ancestry and helps to elaborate strong hypotheses about the functional conservation of BAHD genes in soybean along with those already characterised in different species. The selection history of coding sequences can be assessed through the analysis of base substitutions of the protein code called the K_a/K_s ratio. To investigate the deviation of duplicated BAHD acyltransferases, we determined the non-synonymous substitutions per site (K_a), the synonymous substitutions per site (K_s), and the respective K_a/K_s ratios for each paralogous pair. The partition of each pair was estimated through the K_s values. All paralogous pairs had a K_s value from 0.09 to 0.28, which is consistent with the soybean duplication events (Additional file 2: Table S1). The selection pressure can be determined through the K_a/K_s ratio. Most of the paralogous pairs showed negative Ka/Ks ratios (Additional file 2: Table S1) suggesting that the BAHD gene family in soybean experienced strong purifying selection pressure and was bound by strong evolutionary constraints to maintain its stability [31]. Divergence time analysis showed that these paralogous genes evolved between 7.38 to 22.95 million years ago (Mya),

which is consistent with the recent allotetraploidization of the species that occurred around 13 Mya (Additional file 2: Table S1).

Positive and purifying selection of BAHD genes in soybean

The Ka/Ks ratio is used as a measure of the direction and magnitude of gene selection [38]. The Ka and Ks substitution rates in this study were estimated on the basis of simple probability theory. Positive selection was considered significant with Ka/Ks ratio > 1 [59]. Comparisons of nucleotide substitution per site in the 103 BAHD genes were carried out on the basis of the Ka/Ks ratio in each individual branch of the phylogenetic tree. The neutrality analysis based on maximum likelihood computation of Ka/Ks values of all soybean BAHD genes revealed 16 coding sites under positive selection between codons 69–342 in clade I, 6 coding sites between codons 33–143 in clade II, 19 coding sites in clade III and 15 sites in clade IV under positive selection, all of which showed Ka/Ks ratios > 1 (Additional file 2: Table S2). The changes in codons, as well as sequences for pairwise comparison, were estimated at the Ka and Ks basis. The variability between Ka and Ks polymorphisms at the allelic distribution level presents quite an explicit proof for positive selection [30]. The allelic distribution in various coding sites can be revealed by the strapping test of neutrality, which is applicable to genomic data containing a large number of polymorphisms, along with the homogeneity test allow for comparing the frequency distribution of synonymous sites [3, 46]. The changes in the coding region during the evolutionary time under analysis showed the cumulative presence of synonymous, non-synonymous and ambiguous (indels) codons (Additional file 3: Figure S1). Synonymous and non-synonymous mutations showed low cumulative behaviour at the start site and then gradually increased during the evolutionary time analysed, whereas the opposite behaviour was observed for ambiguous codons at the start point and then acquired a static behaviour at the end of the code. The very same behaviour of synonymous, non-synonymous and ambiguous codons was found in all four clades of the BAHD family (Additional file 3: Figure S1).

We estimated the Likelihood Rate Test (LRT) through the empirical Bayes method at specific branching points and identified various diversifying selection sites. The mixed-effect model of evolution (MEME) was used to identify the diversifying selection of BAHD genes also on the basis of empirical Bayes procedure [85]. MEME identified 7, 17, 16, and 4 episodic diversifying coding sites ($p < 0.05$) in clades I-IV, respectively (Additional file 2: Table S3). The synonymous (α) and non-synonymous (β) substitution rates were calculated, and coding sites with values $\beta > \alpha$ were considered as significant to determine the sites under diversifying selection. The maximum-likelihood estimation of $\beta^{+} > \alpha$ in MEME was obtained for codons 60 and 434 in clade I, 189 and 463 in clade II, 277 and 643 in clade III, and 299 in clade IV (Additional file 2: Table S3). The prevalence of purifying or natural selection masked the episodic natural selection with the transient period of adaptive evolution. Thus, we could not identify sites under positive selection due to the sensitivity of tests used [55] and the stringent parameters chosen, such as $p < 0.05$ and the empirical Bayes factor threshold of analysis for selection of sites (> 15), so that the best outcomes were obtained when the selected branches were placed in the conserved lineage.

Ambiguities in the posterior gene- and site-specific distributions were calculated through FUBAR using the MCMC approach [3, 54]. This analysis revealed the following number of sites in BAHD genes under

pervasive diversifying selection: one site in clade I, four in clades II and III each, and two in clade IV (Additional file 2: Table S4). The diversifying evolutionary sites were detected through the random effect model using the rate distribution with a large number of parameters at the significance level at 95% confidence interval and Pr [β > α] values (Additional file 2: Table S4). The class rate weight at each coding site under positive selection was calculated through a bivariate general discrete distribution. The posterior mean values ranged between 0.80–0.92, which are closer to the value of the potential scale reduction factor (Additional file 2: Table S4) and indicates MCMC convergence. Coding sites with β > α and empirical Bayes factor (EBF) > 15 were considered to be under diversifying selection. The EBF values for each coding site under positive selection were calculated with a net effective sample size (Additional file 2: Table S4). PSRF values for each site under positive selection were < 1.03, whereas the effective population size was > 100. The detected selection across many coding sites was potentially improved through deducing the allocation of gene-specific selection parameters. Altogether, the coding sites under positive selection unveiled in this analysis provide strong evidence of diversifying selection in BADH genes under selection in the soybean lineage.

The evolutionary fingerprint of soybean BAHD genes

Nucleotide and amino acid substitutions in the codon model were used to identify synonymous and nonsynonymous substitutions [4, 48]. The genetic algorithm in the codon model evolution was used to identify the evolutionary fingerprint in coding sites of BAHD genes. The codon model evolution used a phylogenetic Markov model that includes substitution rates, character frequencies [19], amino acid substitution rate clustering [72], and branch lengths estimated through the maximum-likelihood method. The 9052 models generated were used in codon model selection on the basis of likelihood log and modified Bayesian Information Criterion (mBIC). Selective effects with an exchangeable preference for particular amino acid were found in BAHD genes through a model that used the combined empirical codon and transition/transversion-related physiochemical parameters [36]. The model with log (L) value − 16360.3 was considered optimum for the amino acid substitution analysis. We observed mBIC values ranging from 33701.1 for clade II to a maximum value of 81506.1 for clade IV (Additional file 2: Table S5). All clades showed three-class rates for the distribution of amino acids in different classes (Fig. 5) with the estimation of a single Ka/Ks substitution rate (Additional file 2: Table S5). A genetic multi-rate model algorithm was used to analyse the class rates in order to calculate substitution rates at the amino acid level during the evolutionary time scale (Fig. 5). The substitution rate in each class was calculated through genetic models using the Stanfel class parameters (Fig. 5; [72]). The substitution rate distributed the amino acids into the three classes through evolutionary rate cluster, and the substitution pair ACGILMPSTV revealed 90% substitution, whereas DENQ had 50% substitution, and FWY and HKR presented < 50% substitution rates. The best fit model with mBIC values defined the clustering rate, and the numerical values of corresponding rate class substitution are inferred with maximum-likelihood estimation, which were computed evolutionary evidence ratio for the gene [4].

Analysis of cis-regulatory elements in promoter regions of BAHD genes in soybean

Extensive studies on ACTs indicate their roles in plant growth, development, and stress tolerance. Many cis-acting regions are found in the upstream region of GmACTs genes according to our in silico analyses (Additional file 1: File S5). The 103 GmACT promoter regions contain cis-elements responsive to light (43%), hormones (29%), environmental stresses (20%), and plant growth (8%) (Fig. 4A). These regions contained transcription factor binding sites involved in response to plant hormones (Fig. 4B), biotic stresses and pathogen defence, MYB binding sites, and response to heat, low temperature, and drought (Fig. 4C). The pattern of cis-acting regions differed among the soybean BAHD gene promoters (Additional file 1: File S5). GmACT3 (clade I) contains the maximum number of cis-elements (29), whereas GmACT82 (clade IV) contains 26 hormone-responsive elements (Additional file 1: File S5). GmACT34 contains the maximum number of ABA-responsive elements (8), and GmACT45 contains 9 ethylene-responsive motifs (Additional file 1: File S5). GmACT84 and GmACT96 (clade IV), along with GmACT66 (clade III) contain the maximum number of environmental stress-response elements (18, 17, and 15, respectively) in their promoter regions. Members of clades III and IV contain the maximum number of MYB-related cis-elements (Additional file 1: File S5). GmACT84 showed the most (12) MYB related elements, and also showed high expression values, up top seven-fold higher in response to Al^{3+} and two-fold higher to fungal elicitor than the control (Additional file 1: File S1c, S1d). The promoter regions of paralogous pairs were also compared. Both members of each pair showed variation in cis-elements related to stress or hormonal responses or other signals in their promoter regions. Some hormonal, stress or growth-related elements were often present only in one member of the pair (Additional file 1: File S5). Although the GmACT paralogous pairs have little variation in protein length, the sequences of their promoter regions were highly different, which suggests divergent specific roles of each member. We concluded from this analysis that GmACTs may play vital roles in plant stress perception, signalling, or biotic and abiotic stress defence.

Expression patterns of soybean BAHD genes in different tissues

Expression data for the 103 soybean BAHD genes in different soybean tissues were extracted from Phytozome (Fig. 6; Additional file 1: File S1a). Most of the genes showed specific tissue expression patterns related to their clade function. The benzenoid-related genes (GmACT3 and GmACT4) had maximum expression levels in above-ground organs, whereas GmACT5 is highly expressed only in the root (Fig. 6; Additional file 1: File S1a). Benzenoid compounds are produced mostly in flowers and known to be involved in the aroma of numerous fruits and contribute to improving flavour quality [7]. Indeed, already characterised genes from this clade, PhBPBT from Petunia x hybrida and CbBEBT from Clarkia breweri are highly expressed in floral parts and consistently involved with benzylbenzoate compounds [7, 18]. Monolignol synthetic genes (e.g., GmACT1 and GmACT16) showed a high expression level in root and pod. Likewise, OsPMT and BdPMT, which are specific to monolignols with p-coumarate and

ubiquitously expressed but particularly high in the spikelets. When overexpressed, these genes led to a three-fold increase in pCA lignin compared to the control [62]. Monolignol-pCA conjugates integrate with lignin biosynthesis through oxidative radical coupling in order to generate the pCA appendages [62, 84]. Polyamine conjugates are abundant in seeds and serve as nitrogen reserves during germination [23]. The Arabidopsis SDT gene is highly expressed in seeds and produce disinapoyl spermidine and its glucoside while the SCT is mainly expressed in roots and produce coumaroylated spermidines [43]. Soybean also contains SDT and SCT related genes, GmACT2 and GmACT12, respectively, which are expressed in all organs of the plant, although GmACT12 expression is barely detectable in the stem, leaves and roots, although spermidine accumulates at high levels in seeds and the root [43].

The clade II members, GmACT29 and GmACT30, showed high expression only in the seed. GmACT40 is expressed in the root and nodule, while GmACT38 is especially high in nodules and flowers (Fig. 6; Additional file 1: File S1a). On the other hand, GmACT41, GmACT45, GmACT35, GmACT21, GmACT22, and GmACT32 showed high expression levels in all tissues (Fig. 6; Additional file 1: File S1a) suggesting putative roles in cell wall development and plant defence mechanism by producing CGA or quinate-like compounds, given their close relatedness to TpHCT1A and TpHCT1B (Fig. 2; [74]). Phenolamide, an important compound of clade II genes, is often associated with floral induction and development. GmACT21, GmACT22, and GmACT32 showed high transcription levels in the stem and flowers, similarly to TpHCT1, which showed 4 to 5-fold higher expression in these organs than the leaf and played an important role in stem lignification by participating in monolignol biosynthesis [29, 74]. GmACT27, GmACT28 and GmACT45 are similar at the amino acid and also showed the same expression pattern as TpHCT2 [74]. A reduction in the biosynthesis of phenolamides or their conjugates leads to male sterility in plants. Tri-substituted hydroxycinnamoyl spermidines were synthesised in the pollen coat, and this conjugation is catalysed by SHT [21]. GmACT22, GmACT32 and GmACT35 also had their highest expression in the flower and, given their phylogenetic position and expression patterns, these genes might be involved in the synthesis or conjugation of spermidines in the pollen coat of soybean.

Clade III genes, GmACT67 and GmACT72, are highly and specifically expressed in the flower, GmACT55 in the root, GmACT49, GmACT54, and GmACT65 in the nodule, whereas GmACT66 is expressed in the leaf, GmACT47 in stem and flower, GmACT57 in root and flower, and GmACT71 in seed and the pod (Fig. 6; Additional file 1: File S1a). The putative isoflavonoid malonyltransferase genes, such as GmIMaT1, GmMaT2, GmIMaT3, GmMaT4, GmACT68, GmACT69, and GmACT70, showed high expression levels in almost all tissues. Other related members presented a more specific expression pattern, such as GmACT71 in seeds and pods, and GmACT50 in the leaf (Fig. 6; Additional file 1: File S1a). The different expression patterns of these genes suggest putative roles of malonylated flavonoids in storage, defence mechanisms, symbiosis signalling, flower colour, and resistance against biotic and abiotic stresses. Most of clade IIIB members had high expression in leaf and stem, like the characterised members of this group. High and specific expression of CrDAT in the leaf correlates with the synthesis of vindoline from tabersonine [73]. Likewise, the expression of ASAT genes was higher in the trichomes of S. lycopersicum stem and S. pennellii leaf, which correlates with the accumulation of acylsugars in these organs [24].

DCR genes are specifically expressed in the epidermis of vegetative and reproductive organs. Additionally, they also showed expression in the lateral root primordia and developing lateral roots [61]. The cuticle of different organs is the primary barrier against several biotic and abiotic stresses [71]. High expression of DCR genes in young leaves and stem is predictive of their role in the formation of cutin polymers in vegetative organs [61]. During leaf expansion, the cutin polymeric matrix is formed, and DCR plays a vital role to incorporate the hydroxyl fatty acid chains into the cutin polymer [65]. It is also important to convert monomer units into its polymeric form of cutin in the flower. DCR develops the plant lipid polyester for epidermal cell differentiation and patterning, trichome development, and protection from undesired postgenital fusions in aboveground organs [61]. Almost all clade IV genes showed high expression levels in seeds, the root, and nodules. Some genes, such as GmACT82 and GmACT95, presented particularly high expression levels in the root and flowers, GmACT94 and GmACT91 in seeds, GmACT99 in the root and leaves (Fig. 6; Additional file 1: File S1a).

Expression profiles of BAHD genes during nodule development in soybean

The gene expression of BAHD members was analysed in RNA-Seq data derived from soybean nodules at different developing stages [1]. The results showed that BAHD genes were differentially expressed in the nodule at different developmental stages, from nodule development until its senescence (Fig. 7; Additional file 1: File S1b). Most of clade II and clade IV genes showed higher expression in the root than nodules, while 50–80% genes of clades I and III had higher transcription levels in nodules than the root. The expression levels of GmACT39, GmACT40, GmACT41 (clade II), GmACT27 and GmACT33 (clade IV) were 10 to 40-fold higher in the initial stages of nodule development and reached up to > 80-fold in later stages (stage 5, pre-senescence) than in the root. Even though most of the clades I and III genes were expressed more highly in nodules than roots, their expression levels were not much higher than genes in clades II and IV, except for GmACT12 and GmACT68, which showed more than 2 and 70 fold higher expression than root, respectively (Fig. 7; Additional file 1: File S1b). The expression patterns of GmACT35, GmACT39, GmIMaT1, GmACT86, and GmACT87 gradually increased as the nodule developed, peaking at stage 5 and then dropping at senescence (stage 6), which indicates their roles in nodule development. Although the expression levels of GmIMaT1 and GmACT86 were higher in roots than nodules, their expression increased with nodule development. Therefore, these genes may play a symbiotic role in the relationship between root and nodule development by modifying secondary metabolites in roots for secretion into the rhizosphere in order to attract rhizobia, such as isoflavonoids that act as a signalling molecule for inducing nod factor biosynthesis in rhizobia. The transcription level of GmIMaT3 was the highest after GmACT87 than all other BAHD genes in the different nodule stages analysed, and its expression remained static throughout the nodule development and almost similar with that in the root (Fig. 7; Additional file 1: File S1b). The expression levels of GmACT15, GmACT18, GmACT40, GmACT41, GmMaT2, and GmACT75 were higher in the initial stages and gradually decreased as the nodule development advanced, reaching the lowest expression levels at its final stage (Fig. 7; Additional file 1: File S1b). Therefore, these genes may play a function in rhizobial interaction and nodule initiation and development.

Expression patterns of soybean BAHD genes in response to different stresses

The expression of BAHD family genes was analysed from RNA-Seq data obtained from plants subjected to acidic condition (low pH 4.0), aluminium treatment 50 μM Al^{3+} (pH 4.0), and fungal elicitor after 2 and 24 h. Most of the BAHD genes expression decreased in response to low pH and Al^{3+}, except for GmACT13 (clade I) and GmACT36 (clade II). Their expression levels doubled at low pH but decreased or remained the same at Al^{3+} stress conditions (Fig. 8; Additional file 1: File S1c). The expression of GmACT49, GmIMaT1, GmIMaT3 (clade III) and GmACT84 and GmACT92 (clade IV) increased dramatically in response to both stress conditions (Fig. 8; Additional file 1: File S1c). The expression of GmACT49, GmACT84 and GmACT92 increased about 11, 7, 40 fold in response to low pH, and 6, 7, 70 fold in response to Al^{3+} stress, respectively, whereas the expression of GmIMaT1 and GmIMaT3 significantly increased as compared to the control under these conditions (Fig. 8; Additional file 1: File S1c). Therefore, the BAHD genes that showed elevated transcription levels after low pH and Al^{3+} treatments could be good candidate genes for soybean to Al^{3+} toxicity tolerance. The transcription levels of GmACT12 (clade I), GmACT39 (clade II), GmACT62, GmACT68, GmMaT2, GmMaT4, GmACT75 (clade III) and GmACT86 (clade IV) decreased from 5 to 70-fold, revealing their sensitivity towards low pH and Al^{3+} stress conditions.

Transcriptomic analyses in response to fungal elicitor (FE) showed that the expression of 29% of the BAHD genes (30 out of the 103) changed significantly, with 10 genes increasing and 20 decreasing their expression levels (Fig. 9; Additional file 1: File S1d). The expression of GmACT35, GmACT36 (clade II) GmIMaT1, GmMaT2, GmACT50, GmACT57, GmACT58 (clade III), GmACT84, GmACT85 and GmACT92 (clade IV) increased while the other genes had their transcription level decreased. The expression of GmACT36, GmACT57, GmIMaT1, GmACT58, GmACT84 and GmACT92 increased almost 12, 20, 2, 40, 2, and 2 folds, respectively (Fig. 9; Additional file 1: File S1d). This expression pattern suggests a role in resistance to fungal attack by the synthesis and secretion of modified metabolites. Indeed, fungal and bacterial infections, or elicitor treatment in Solanaceae plants and Arabidopsis, increase the synthesis of coumaroyl or feruloyl tyramines [53, 87]. The anti-microbial activity was demonstrated for dicoumaroyl-caffeoyl spermidine, which inhibits mycelial growth of Pyrenophora and reduces powdery mildew infection in barley seedlings [83]. Likewise, constitutive accumulation of tyramine derivatives in transgenic tomato promotes resistance to Pseudomonas syringae [11]. MtMaT3 expression increased in response to fungal infection and decreased in response to copper treatment [86]. The observed modulation of BAHD family genes in soybean in response to fungal elicitor may correlate with an increased acylation of active compounds against fungal attacks.

Expression patterns of soybean BAHD family genes in different soybean genotypes seed coat and cotyledons

Soybean Illumina expression data of seed coat and cotyledon obtained from developing seed (stage 5, 380–450 mg in weight) of three varieties were downloaded [13] and analysed. Interestingly, some BAHD family genes were highly expressed in seed coat while expressed lowly in cotyledons, and vice versa. GmACT5 (clade I), GmACT25 (clade II), GmMaT2 and GmIMaT3 (clade III), and GmACT84 (clade IV) showed high expression in the cotyledon as compared to seed coat (Additional file 1: File S1e) suggesting a function in the storage of modified metabolites, or resistance against seed borers. The expression levels of GmACT4, GmACT15 (clade I), GmACT21, GmACT22, and GmACT35 (clade II), GmACT57, GmMaT4, GmACT70 (clade III), and GmACT79, GmACT81 (clade IV) were more specific and high in brown than blackl seed coats and almost absent in yellow seed coats (Additional file 1: File S1e) suggesting a role in anthocyanin modification.

Expression modulation of soybean BAHD family genes by MYB transcription factors, GmLEC2a and GmWRi1b in the hairy root heterologous system

MYB transcription factors (TF) tightly regulate the biosynthesis of flavonoids and anthocyanins [41]. Many soybean MYB factors have been reported to either positively or negatively regulate isoflavonoid accumulation in roots and seeds or the production of anthocyanins and proanthocyanidins (PAs) in seeds [41]. Two soybean MYB TFs, including one CPC-like R3-MYB repressor (GmMYB3), and an R2R3-MYB activator (GmMYB7), were over-expressed in soybean hairy roots and an RNA-Seq analysis was performed. The expression of GmACT17 (clade I), GmACT23, GmACT40, GmACT41, GmACT43 (clade II), GmACT75 (clade III) and GmACT98 (clade IV) was upregulated by GmMYB3 (Additional file 1: File S1f; Additional file 3: Fig. S2) while the expression of GmACT9, GmACT12 (clade I), GmACT21, GmACT22, GmACT28 (clade II) was induced by GmMYB7 (Additional file 1: File S1f; Additional file 3: Fig. S2). The transcription levels of GmACT24 (clade II) and GmMaT2 (clade III) were 2 to 5-fold upregulated by both, GmMYB3 and GmMYB7, as compared to the control. Most of the induced genes belong to clades I and II, which are involved in the modification of different compounds that by adding ferulyl, caffeoyl, sinopoyl, p-coumaroyl, and hydroaxycinnamoyl moieties. On the other hand, the upregulated genes of clade III may be involved in the synthesis of flavonoids or anthocyanins since we observed the over-expression of MtPAR, which a previous report showed to lead a 10-fold increase in the flavonoid contents [41]. GmMYB3 and GmMYB7 negatively regulated most of the clade III genes, but not GmACT44 (clade II). The expression levels of GmACT54, GmACT57, GmACT62, GmACT64 and GmACT65 were suppressed in the MYB over-expressed soybean hairy roots (Additional file 1: File S1f; Additional file 3: Fig. S2). High isoflavonoid levels were reported in soybean hairy roots [28]. MtPAR in soybean hairy roots decreased contents of isoflavonoids, such as daidzin, glycitin, genistin, malonyldaidzin, and daidzein up to 3, 2, 4, 2.5, and 8-fold, respectively, as compared to control through repressing the IFS gene [42]. In fact, MYB-TF suppressed the expression of not only IFS but also other isoflavonoid biosynthesis genes. An active complex is formed to activate isoflavone biosynthesis via a MYB activator [42]. MtPAR also has a negative impact on isoflavone production through expression suppression of the biosynthesis genes [41].

Some of the BAHD acyltransferases were also significant up- or down-regulated in GmLEC2a-over-expressing hairy roots [47]. While some genes specifically expressed in leaves (GmACT26) and roots (GmACT84, GmACT85) should not be native targets of GmLEC2a, genes that are highly expressed in seeds (GmACT15, GmACT34), pod (GmACT71, GmACT99) or open flowers (GmACT75, GmACT97) may be natural GmLEC2a targets. Many of such genes were repressed in the GmLEC2a-overexpressing system (Additional file 1: File S1g), suggesting that LEC2a may mainly upregulate the primary metabolism while repressing specialised metabolic genes, including those coding for enzymes that modify metabolic structures, such as BAHD genes.

Some of these secondary metabolic genes were also regulated by GmWRI1b, as shown by their up- or downregulation in GmWRI1b-OE transgenic hairy roots [13]. There were 63% (65 out of 103 genes) BAHD genes with significantly modulated expression in GmWRI1b-OE hairy roots, being 51% of them upregulated and 47% downregulated (Additional file 1: File S1h). Interestingly, some genes with high expression in aboveground organs (e.g., GmACT2, GmACT4) were upregulated in GmWRi1b-OE hairy roots. BAHD genes that are expressed specifically in leaves (GmACT28), stems (GmACT47), and flowers (GmACT47, GmACT57, GmACT82) were upregulated in GmWRi1b-OE hairy roots. Nodule expressed genes (GmACT9, GmACT44, GmACT49) were upregulated whereas, GmACT54, GmACT87 and GmACT89 were downregulated. On the other hand, most of the genes specifically expressed in roots (GmACT18, GmACT21, GmACT22, GmACT23, GmACT36, GmACT40, GmACT41, GmACT43, GmACT86) and seeds (GmACT15, GmACT94) were downregulated in GmWRi1b-OE hairy roots. GmACT56 and GmACT62 were highly expressed in roots along with other organs were upregulated but those root-expressed genes that were also expressed in nodules (e.g., GmACT92, GmACT98) were down-regulated. We also noticed that some of BAHD members with very low count values in different soybean organs were significantly upregulated (GmACT33) or downregulated (GmACT17) in the GmWRi1b-overexpressing hairy root system. Most of the clade IV genes that targeted acyl-lipids for wax or cuticle synthesis were down-regulated in the heterologous root system (Additional file 1: File S1h).

qRT-PCR validation of BAHD genes in developing nodules and in response to stresses

In order to validate in silico expression and further reveal the involvement of BAHD genes in physiological processes, we performed quantitative reverse-transcription PCR (qRT-PCR). We analysed the expression profiles during nodule development and the response to Al^{3+} and fungal elicitor exposures. To select GmBAHD genes for expression analyses, we considered the following criteria: (i) high expression in different tissues; (ii) representation of genes in all clades; (iii) inclusion of some paralogous pairs. The expression profiles revealed by qRT-PCR were generally consistent with the RNA-seq data. The expression of GmACT2 was higher in nodules than roots but static in all nodule developing stages, while the expression levels of GmACT21, GmACT22, GmACT55, GmACT68, GmACT79, GmACT81, GmACT84 and GmACT86 were higher in roots while lower but static during nodule development. GmACT12, GmACT15, GmACT41 expression was higher during the initial stages of nodule development but gradually decreased

and remained low during nodule senescence. GmACT44 expression was initially low and gradually increased and was found high expression during nodule senescence (Fig. 7B). GmACT84 expression was high in response to Al^{3+} (Fig. 8B) and fungal elicitor (Fig. 9B) compared to the control while most of the other genes tested decreased their expression significantly or remained almost unchanged.

Conclusions

The findings described here will help reveal the roles that BAHD genes play in soybean as well as in other plant species. We identified 103 BAHD genes in the soybean genome spread over all its 20 chromosomes. The distribution of genes within the genome, their organization and structure suggest a complex evolutionary history of the BAHD family in this legume species. Several cases of segmentally duplicated genes were found, which could have originated during the genome duplication event. The expression dynamics of BAHD genes in different tissues of the mature plant, during nodule development, and in response to different stressors revealed that this family is broadly involved in soybean organogenesis and stress responses, which warrant further investigation on the molecular and physiological functions of members of this family.

Methods

Identification of BAHD family genes in the Glycine max genome

BLASTP against the G. max proteome dataset obtained from the genome sequence (https://phytozome.jgi.doe.gov/pz/portal.html) was used to search for BAHD-coding genes using two characterised protein sequences from legumes as query: GmMT7 from soybean [20], and MaTs from Medicago truncatula [86, 89]. The HXXXD motif was searched manually in each retrieved sequence to confirm it as a candidate BAHD protein. Features of BAHD proteins, such as the number of amino acid residues, along with the start-to-end position of their respective gene in the genome, and chromosome location were acquired from the Phytozome database. Physical protein parameters of each putative gene product, such as molecular weight (Mw) and isoelectric point (pI) were calculated from pI/Mw tool at ExPASy (http://www.expasy.org/tools/) using the default parameters.

Chromosomal location, phylogeny, and gene structure analysis of the BAHD genes

PhenoGram Plot (http://visualization.ritchielab.psu.edu/phenograms/plot) was used to create the image of chromosomal locations of BAHD family genes using the chromosomal position information available at Phytozome. A phylogenetic analysis with functionally characterised proteins from different species and all BAHD proteins encoded in the soybean genome was conducted to explore the evolutionary

relationships of this gene family. An unrooted phylogenetic tree was constructed following the Neighbour-Joining method in MEGA 6.0 [78]. The tool Gene Structures Display Server (http://gsds.cbi.pku.edu.cn) was used to determine the exon/intron structures of individual BAHD genes.

Calculation of K_a and K_s ratios

The Plant Genome Duplication Database (PGDD, http://chibba.agtec.uga.edu/duplication/) was used to obtain the K_a and K_s values of BAHD paralogous pairs. Divergence time (T) was calculated using the formula $T = K_s/2\lambda$, where $\lambda = 6.161029 \times 10^{-9}$ for G. max [45]. Selection pressures on duplicated genes were estimated using the K_a/K_s ratio for each paralogous gene pair. Assumptions of negative, neutral and positive evolution were considered for the values < 1, =1, and > 1, respectively.

Diversifying evolutionary selection analysis

A phylogenetic analysis was carried out by the HyPhy program on cDNA sequences using the Nei-Gojobori method and applied in MEGA 6.0 to calculate K_a/K_s ratios for neutrality analysis of BAHD genes within each clade. Maximum-likelihood based on the Kimura-2 parameter model was used to infer the evolutionary history of the genes within each clade. The detection of episodic diversification effect of individual coding sites was performed through different approaches. The episodic diversifying selection and pervasive positive selection were identified at the individual branch site level through the mixed-effect model of evolution (MEME). Markov Chain Monte Carlo (MCMC) was used in fast unconstrained Bayesian approximation (FUBAR) to ensure the strength over predefined sites through approximate Bayesian method [55]. The parameter $\omega = \beta/\alpha$ was used in MEME to fit the data in the GTR nucleotide model as initial values. The $\beta{:}\beta^- \leq \alpha$ and β^+ parameters were used to measure the selection pressure, whereas β^-, β^+ and α were used to estimate the variability of site-to-site rate substitution. LRT based on χ^2 asymptotic distribution was considered significant at p-value < 0.05.

Identification of conserved motifs and promoter region analysis

The analysis of conserved region within BAHD family genes was executed through MEME (http://meme-suite.org/), and their genomic assemblies were screened through the Pfam database (http://pfam.sanger.ac.uk). Regulatory elements and their analytical data were obtained using the Plant cis-acting Regulatory DNA Elements (PlantCARE) program (http://bioinformatics.psb.ugent.be/webtools/plantcare) to analyse the 1.5 kb promoter region upstream of the start codon for each BAHD gene.

Gene expression analysis

Gene expression data derived from microarrays of identified BAHD genes in seven tissues (seed, root, nodule, stem, leaf, flower, and pod; Additional file 1: File S1a) was obtained at Phytozome v.12.0 (https://phytozome.jgi.doe.gov/pz/portal.html).

Plant growth, RNA-Seq data analysis, and gene expression confirmation with qRT-PCR

Soybean seeds cv. Williams 82 were germinated in vermiculite in a light chamber at 25 ± 2 ºC. Samples at different nodulation stages (N-1 to N-6) were obtained [1]. The seedlings were hydroponically subjected to acidic treatment (pH 4.0) and 50 mM Al^{3+} stress (under pH 4.0) [2, 82] for 10 days before harvesting the roots for gene expression analysis.

The samples were collected, snap frozen in liquid N_2, and stored at − 80 ºC for RNA extraction. Solexa sequencing libraries were synthesised and further sequenced. Data were utilised to quantify the expression of soybean BAHD family genes (i.e. the number of sequence reads/million reads aligned) in nodule developing stages (Additional files 1: File S1b), response in Al^{3+} stressed roots (Additional files 1: File S1c) and fungal elicitor treated roots (Additional files 1: File S1d), genotypes with different seed coat colour (Additional files 1: File S1e), as well as hairy roots overexpression the following genes: MYB transcription factors (MYB3 and MYB7) (Additional files 1: File S1f), GmLEC2a (Additional files 1: File S1g), and GmWRi1b (Additional files 1: File S1h).

Fourteen BAHD genes were selected on the basis of their expression level in different soybean tissues for validation of their expression during nodule developing and response to Al^{3+} and fungal elicitor through qRT-PCR with gene-specific primers (Additional files 1: File S2). An iQ5 Real-Time PCR System (Bio-Rad) with 96-well plates with each a reaction volume of 20 µL was used. Each reaction consisted of 2.5 µL of Power SYBR Master Mix (Applied Biosystems), 1 µL primer mix (0.4 µL of each F/R primer + 0.2 µL H_2O) and 2 µL cDNA. Gene expression was normalised with the housekeeping gene GmACTIN.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the National Science Foundation of China (grant 31670294), the Ministry of Science and Technology of China (grant 2016YFD0100504, 2016YFD0101005), and the funding from Anhui Agricultural University and the State Key Laboratory of Tea Plant Biology and Utilization.

Authors' contributions

JZ planned and designed the research. MZ, XZ, QD, SM, PL, XW performed experiments and bioinformatics analyses. JZ and MZ wrote the manuscript. VAB proofread and edited the manuscript. All authors have read and approved the manuscript.

Acknowledgements

The authors thank lab members in Prof. Zhao’s lab for all assistances in experiments and data analyses.

Abbreviations

ACT: Acyltransferase

UV: Ultraviolet

HCBT: Hydroxycinnamoyl/benzoyl-CoA:anthranilate transferase

HCT: Hydroxycinnamoyl transferase

SHT: Spermidine hydroxycinnamoy ltransferase

HCA: Hydroxycinnamic acid

MaT: Malonyltransferase

MW: Molecular weight

pI: Isoelectric point

cDNA: Complementary deoxynucleic acid

PlantCARE: Plant cis-acting Regulatory DNA Elements

MEME: Mixed-effect model evolution

MCMC: Markov Chain Monte Carlo

FUBAR: Fast unconstrained Bayesian approximation

kDs: Kilodaltons

WGD: Whole-genome duplication

MYA: Million years ago

EBF: Empirical Bayes factor

mBIC: Modified Bayesian Information Criterion

References

1. Ahmad MZ, Rehman NU, Yu S, Zhou Y, Haq BU, Wang J, Li P, Zeng Z, Zhao J. GmMAX2–D14 and –KAI interaction-mediated SL and KAR signaling play essential roles in soybean root nodulation. The Plant Journal. 2019;Doi: 10.1111/tpj.14545.
2. Ahmad MZ, Li P, Wang J, Rehman NU, Zhao J. Isoflavone malonyltransferases *GmIMaT1* and *GmIMaT3* differently modify isoflavone glucosides in soybean (*Glycine max*) under various stresses. Frontiers in Plant Science. 2017;8:735.
3. Ahmad HI, Liu G, Jiang X, Liu C, Xu F, Chong Y, Ijaz N, Huang H. Adaptive selection at agouti signaling protein gene inferred breed specific selection signature within the indigenous goat populations. Asian-Australasian Journal of Animal Sciences. 2017a;1-7. Doi.org/10.5713/ajas.16.0994.
4. Ahmad HI, Liu G, Jiang X, Liu C, Chong Y, Huang H. Adaptive molecular evolution of *MC1R* gene reveals the evidence for positive diversifying selection in indigenous goat populations. Ecology and Evolution. 2017b;DOI:10.1002/ece3.2919.
5. Bartley LE, Peck ML, Kim S-R, Ebert B, Manisseri C, Chiniquy DM, Sykes R, Gao L, Rautengarten C, Vega-Sanchez ME, et al. Overexpression of a BAHD acyltransferase, *OsAt10*, alters rice cell wall hydroxycinnamic acid content and saccharification. Plant Physiology. 2013;161:1615–1633.
6. Besseau S, Hoffmann L, Geoffroy P, Lapierre C, Pollet B, Legrand M. Flavonoid accumulation in *Arabidopsis* repressed in lignin synthesis affects auxin transport and plant growth. Plant Cell. 2007;19:148– 162.
7. Boatright J, Negre F, Chen X, Kish CM, Wood B, Peel G, Orlova I, Gang D, Rhodes D, Dudareva N. Understanding in vivo benzenoid metabolism in petunia petal tissue. Plant Physiology. 2004;135:1993–2011.
8. Bloch SE, Schmidt-Dannert C. Construction of a chimeric biosynthetic pathway for the de novo biosynthesis of rosmarinic acid in *Escherichia coli*. ChemBiochem. 2014;15:2393–401.
9. Bontpart T, Cheynier V, Ageorges A, Terrier N. BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. New Phytologist. 2015;208:695–707.
10. Bowles D, Isayenkova J, Lim EK, Poppenberger B. Glycosyltransferases: managers of small molecules. Current Opinion in Plant Biology. 2005;8:254–263.
11. Campos L, Lisón P, López-gresa MP, Rodrigo I, Zacarés L, Conejero V, Bellés JM. Transgenic tomato plants overexpressing tyramine N-hydroxycinnamoyltransferase exhibit elevated hydroxycinnamic

acid amide levels and enhanced resistance to *Pseudomonas syringae*. Molecular Plant-Microbe Interactions. 2014;27:1159–1169.

12. Chedgy RJ, K€ollner TG, Constabel CP. Functional characterization of two acyltransferases from *Populus trichocarpa* capable of synthesizing benzyl benzoate and salicyl benzoate, potential intermediates in salicinoid phenolic glycoside biosynthesis. 2015;113:149–159.
13. Chen B, Zhang G, Li P, Yang J, Guo L, Benning C, Wang X, Zhao J. Multiple *GmWRI1s* are redundantly involved in seed filling and nodulation by regulating plastidic glycolysis, lipid biosynthesis and hormone signalling in soybean (*Glycine max*). Plant Biotechnology Journal. 2020; 18(1):155-171
14. Cheynier V, Comte G, Davies KM, Lattanzio V, Martens S. Plant phenolics: recent advances on their biosynthesis, genetics, and ecophysiology. Plant Physiology and Biochemistry. 2013;72:1–20.
15. D'Auria JC, Reichelt M, Luck K, Svatos A, Gershenzon J. Identification and characterization of the BAHD acyltransferase malonyl CoA:anthocyanidin 5-*O* glucoside-6-O malonyltransferase (*At5MAT*) in *Arabidopsis thaliana*. FEBS Letters. 2007;581:872–878.
16. D'Auria JC. Acyltransferases in plants: a good time to be BAHD. Current Opinion in Plant Biology. 2006;9:331–340.
17. D'Auria JC, Gershenzon J. The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. Current Opinion in Plant Biology. 2005;8:308–316.
18. D'Auria JC, Chen F, Pichersky E. Characterization of an acyltransferase capable of synthesizing benzylbenzoate and other volatile esters in flowers and damaged leaves of *Clarkia breweri*. Plant Physiology. 2002;130:466–476.
19. Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, Kosakovsky Pond SL. CodonTest: modeling amino acid substitution preferences in coding sequences. PLOS Computational Biology. 2010;6(8). pii: e1000885.
20. Dhaubhadel S, Farhangkhoee M, Chapman R. Identification and characterization of isoflavonoid specific glycosyltransferase and malonyltransferase from soybean seeds. Journal of Experimental Botany. 2008;59:981–994.
21. Elejalde-Palmett C, de Bernonville TD, Glevarec G, Pichon O, Papon N, Courdavault V, St-Pierre B, Giglioli-Guivarc'h N, Lanoue A, Besseau S. Characterization of a spermidine hydroxycinnamoyltransferase in *Malus domestica* highlights the evolutionary conservation of trihydroxycinnamoyl spermidines in pollen coat of core Eudicotyledons. Journal of Experimental Botany. 2015; doi:10.1093/jxb/erv423
22. Eudes A, Juminaga D, Baidoo EE, Collins FW, Keasling JD, Loqué D. Production of hydroxycinnamoyl anthranilates from glucose in *Escherichia coli*. Microbial Cell Factories. 2013;12:62.
23. Facchini PJ, Hagel J, Zulak KG. Hydroxycinnamic acid amide metabolism: physiology and biochemistry. Canadian Journal of Botany. 2002;80:577–589.
24. Fan P, Miller AM, Schilmiller AL, Liu X, Ofner I, Jones AD, Zamir D, Last RL. In vitro reconstruction and analysis of evolutionary variation of the tomato acylsucrose metabolic network. Proceedings of the National Academy of Sciences, USA. 2015;113(2):E239-48.

25. Fujiwara H, Tanaka Y, Yonekura-Sakakibara K, Fukuchi-Mizutani M, Nakao M, Fukui Y, Yamaguchi M, Ashikari T, Kusumi T. cDNA cloning, gene expression and subcellular localization of anthocyanin 5-aromatic acyltransferase from *Gentiana triflora*. Plant Journal. 1998;16:421–431.
26. Gomez C, Terrier N, Torregrosa L, Vialet S, Fournier-Level A, Verries C, Souquet JM, Mazauric JP, Klein M, Cheynier V, et al. Grapevine MATE-type proteins act as vacuolar H^+-dependent acylated anthocyanin transporters. Plant Physiology. 2009;150:402–415.
27. Grienenberger E, Besseau S, Geoffroy P, Debayle D, Heintz D, Lapierre C, Pollet B, Heitz T, Legrand M. A BAHD acyltransferase is expressed in the tapetum of *Arabidopsis* anthers and is involved in the synthesis of hydroxycinnamoyl spermidines. Plant Journal. 2009;58:246–259.
28. Gutierrez-Gonzalez JJ, Guttikonda SK, Tran LS, Aldrich DL, Zhong R, Yu O, Nguyen HT, et al. Differential expression of isoflavone biosynthetic genes in soybean during water deficits. Plant and Cell Physiology. 2010;51:936–948.
29. Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, Pollet B, Legrand M. Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. Plant Cell. 2004;16:1446–1465.
30. Hughes AL, Friedman R. Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. Immunogenetics. 2008;60:495-506.
31. Juretic N1, Hoen DR, Huynh ML, Harrison PM, Bureau TE. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Research. 2005;15:1292–1297.
32. Kannangara R, Branigan C, Liu Y, Penfield T, Rao V, Mouille G, Höfte H, Pauly M, Riechmann JL, Broun P. The transcription factor WIN1/SHN1 regulates cutin biosynthesis in *Arabidopsis thaliana.* Plant Cell. 2007;19:1278–1294.
33. Kim BG, Jung WD, Mok H, Ahn JH. Production of hydroxycinnamoyl-shikimates and chlorogenic acid in *Escherichia coli:* production of hydroxycinnamic acid conjugates. Microbial Cell Factories. 2013;12:15.
34. Kim IA, Kim BG, Kim M, Ahn JH. Characterization of hydroxycinnamoyltransferase from rice and its application for biological synthesis of hydroxycinnamoyl glycerols. Phytochemistry. 2012;76:25–31.
35. Kim J, Kang K, Gonzales-Vigil E, Shi F, Jones AD, Barry CS, Last RL. Striking natural diversity in glandular trichome acylsugar composition is shaped by variation at the Acyltransferase2 locus in the wild tomato *Solanum habrochaites*. Plant Physiology. 2012;160(4):1854-70.
36. Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. Molecular Biology and Evolution. 2007;24:1464–1479.
37. Kosma DK, Molina I, Ohlrogge JB, Pollard M. Identification of an *Arabidopsis* fatty alcohol:caffeoyl-coenzyme A acyltransferase required for the synthesis of alkyl hydroxycinnamates in root waxes. Plant Physiology. 2012;160:237–248.
38. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genetics. 2008;4:e1000304.

39. Lam KC, Ibrahim RK, Behdad B, Dayanandan S. Structure, function, and evolution of plant O-methyltransferases. Genome. 2007;50:1001–1013.
40. Landry LG, Chapple CCS, Last RL. *Arabidopsis* mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage. Plant Physiology. 1995;109:1159–1166.
41. Li P, Dong Q, Ge S, He X, Verdier J, Li D, Zhao J. Metabolic engineering of proanthocyanidin production by repressing the isoflavone pathways and redirecting anthocyanidin precursor flux in legume. Plant Biotechnology Journal. 14(7):1604-18.
42. Li X, Dhaubhadel S. 14-3-3 proteins act as scaffolds for GmMYB62 and GmMYB176 and regulate their intracellular localization in soybean. Plant Signaling & Behavior. 2012;7:965–968.
43. Luo J, Fuell C, Parr A, Hill L, Bailey P, Elliott K, Fairhurst SA, Martin C, Michael AJ. A novel polyamine acyltransferase responsible for the accumulation of spermidine conjugates in *Arabidopsis* Plant Cell. 2009;21:318–333.
44. Luo J, Nishiyama Y, Fuell C, et al. Convergent evolution in the BAHD family of acyl transferases: identification and characterization of anthocyanin acyl transferases from *Arabidopsis thaliana*. Plant Journal. 2007;50:678–695.
45. Lynch M, Conery JS. The Evolutionary Fate and Consequences of Duplicate Genes. Science. 2000;290:1151.
46. Manel S, Perrier C, Pratlong M, et al. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. Molecular Ecology. 2016;25:170-84.
47. Manan S, Ahmad MZ, Zhang G, Chen B, Haq BU, Yang J, Zhao J. Soybean LEC2 regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. Frontiers in Plant Science. 2017;8:1604.
48. Miyazawa S. Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. PLoS ONE. 2011;6:e28892.
49. Moglia A, Acquadro A, Eljounaidi K, Milani AM, Cagliero C, Rubiolo P, Genre A, Cankar K, Beekwilder J, Comino C. Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in caffeoylquinic acid synthesis in *Globe Artichoke*. Frontiers in Plant Science. 2016;7:1424.
50. Molina I, Li-Beisson Y, Beisson F, Ohlrogge JB, Pollard M. Identification of an *Arabidopsis* feruloyl-coenzyme A transferase required for suberin synthesis. Plant Physiology. 2009;151:1317–1328.
51. Mondolot L, La Fisca P, Buatois B, Talansier E, de Kochko A, Campa C. Evolution in caffeoylquinic acid content and histolocalization during Coffea canephora leaf development. Annals of Botany. 2006;98(1):33-40.
52. Mugford ST, Qi X, Bakht S, et al. A Serine Carboxypeptidase-Like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. Plant Cell. 2009;21:2473–2484.
53. Muroi A, Ishihara A, Tanaka C, Ishizuka A, Takabayashi J, Miyoshi H and Nishioka T. Accumulation of hydroxycinnamic acid amides induced by pathogen infection and identification of agmatine

coumaroyltransferase in *Arabidopsis thaliana*. Planta. 2009;230:517–527.

54. Murrell B, Moola S, Mabona A, et al. FUBAR: A fast, unconstrained Bayesian approximation for inferring selection. Molecular Biology and Evolution. 2013;30:1196–1205.
55. Murrell B, Wertheim JO, Moola S, et al. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics. 2012;8:e1002764.
56. Nakayama T, Suzuki H, Nishino T. Anthocyanin acyltransferases: specificities, mechanism, phylogenetics, and applications. Journal of Molecular Catalysis B: Enzymatic. 2003;23:117–132.
57. Narukawa M, Kanbara K, Tominaga Y, et al. Chlorogenic acid facilitates root hair formation in lettuce seedlings. Plant & Cell Physiology. 2009;50(3):504-514.
58. Niggeweg R, Michael AJ, Martin C. Engineering plants with increased levels of the antioxidant chlorogenic acid. Nature Biotechnology. 2004;22:746–754.
59. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365:185-205.
60. Onkokesung N, Gaquerel E, Kotkar H, Kaur H, Baldwin IT, Galis I. MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A:polyamine transferases in *Nicotiana attenuata*. Plant Physiology. 2012;158:389–407.
61. Panikashvili D1, Shi JX, Schreiber L, Aharoni A. The *Arabidopsis* DCR encoding a soluble BAHD acyltransferase is required for cutin polyester formation and seed hydration properties. Plant Physiology. 2009;151(4):1773-89.
62. Petrik DL, Karlen SD, Cass CL, et al. p-Coumaroyl-CoA:monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in *Brachypodium distachyon*. Plant Journal. 2014;77:713–726.
63. Pollard M, Beisson F, Li Y, Ohlrogge JB. Building lipid barriers: biosynthesis of cutin and suberin. Trends in Plant Science. 2008;13:236–246.
64. Rautengarten C, Ebert B, Ouellet M, et al. Arabidopsis deficient in cutin ferulate encodes a transferase required for feruloylation of x-hydroxy fatty acids in cutin polyester. Plant Physiology. 2012;158:654–665.
65. Reina JJ, Guerrero C, Heredia A. Isolation, characterization, and localization of *AgaSGNH* cDNA: a new SGNH-motif plant hydrolase specific to *Agave americana* leaf epidermis. Journal of Experimental Botany. 2007;58:2717–2731.
66. Santos ALD, Chaves-Silva S, Yang L, Maia LGS, Chalfun-Júnior A, Sinharoy S, Zhao J, Benedito VA. Global analysis of the MATE gene family of metabolite transporters in tomato. BMC Plant Biology. 2017;17(1):185.
67. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–183.
68. See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, Gill BS. Gene evolution at the ends of wheat chromosomes. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(11):4162–7.

69. Serra O, Hohn C, Franke R, Prat S, Molinas M, Figueras M. A feruloyl transferase involved in the biosynthesis of suberin and suberin-associated wax is required for maturation and sealing properties of potato periderm. Plant Journal. 2010;62:277–290.
70. Severin AJ, Woody JL, Bolon Y-T, et al. RNASeq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biology. 2010;10(1):160.
71. Shepherd T, Wynne Griffiths D. The effects of stress on plant cuticular waxes. New Phytologist. 2006;171:469–499.
72. Stanfel LE. A new approach to clustering the amino acid. Journal of Theoretical Biology. 1996;183;195–205.
73. St-Pierre B, Laflamme P, Alarco AM, De Luca V. The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. the Plant Journal. 1998;14(6):703-13.
74. Sullivan ML. A novel red clover hydroxycinnamoyl transferase has enzymatic activities consistent with a role in phaselic acid biosynthesis. Plant Physiology. 2009;150:1866–1879.
75. Suzuki H, Nishino T, Nakayama T. cDNA cloning of a BAHD acyltransferase from soybean (*Glycine max*): isoflavone 7-*O* glucoside- 6-*O* Phytochemistry. 2007;68:2035–2042.
76. Suzuki H, Nakayama T, Nagae S, Yamaguchi M, Iwashita T, Fukui Y, Nishino T. cDNA cloning and functional characterization of flavonol 3-*O*-glucoside 6-*O*-malonyltransferases from flowers of *Verbena hybrida* and *Lamium purpureum*. Journal of Molecular Catalysis B: Enzymatic. 2004;28:87–93.
77. Suzuki H, Nakayama T, Yonekura-Sakakibara K, Fukui Y, Nakamura N, Nakao M, Tanaka Y, Yamaguchi M, Kusumi T, Nishino T. Malonyl-CoA: anthocyanin 5-O-glucoside-6"-O-malonyltransferase from scarlet sage (*Salvia splendens*) flowers. Journal of Biological Chemistry. 2001;276:49013–49019.
78. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution. 2013;30(12):2725–9.
79. Tremblay Y. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biology. 2004;4:10.
80. Varbanova M, Porter K, Lu F, Ralph J, Hammerschmidt R, Jones AD, Day B. Molecular and biochemical basis for stress-induced accumulation of free and bound p-coumaraldehyde in cucumber. Plant Physiology. 2011;157(3):1056-66.
81. Wang J, De Luca V. The biosynthesis and regulation of biosynthesis of Concord grape fruit esters, including ‘foxy’methylanthranilate. Plant Journal. 2005;44:606–619.
82. Wang J, Hou Q, Li P, Yang L, Sun X, Benedito VA, Wen J, Chen B, Mysore KS, Zhao J. Diverse functions of multidrug and toxin extrusion (MATE) transporters in citric acid efflux and metal homeostasis in *Medicago truncatula*. Plant Journal. 2017;90:79–95.
83. Walters D, Meurer-Grimes B, Rovira I. Antifungal activity of three spermidine conjugates. FEMS Microbiology Letters. 2001;201:255–258.

84. Withers S, Lu F, Kim H, Zhu Y, Ralph J, Wilkerson CG. Identification of grass-specific enzyme that acylates monolignols with p-coumarate. Journal of Biological Chemistry. 2012;287:8347–8355.
85. Yang Z, Dos Reis M. Statistical properties of the branch-site test of positive selection. Molecular Biology and Evolution. 2011;28:1217–1228.
86. Yu XH, Chen MH, Liu CJ. Nucleocytoplasmic-localized acyltransferases catalyze the malonylation of 7-O-glycosidic (iso)flavones in *Medicago truncatula*. the Plant Journal. 2008;55:382–396.
87. Zacarés L, López-Gresa MP, Fayos J, Primo J, Bellés JM, Conejero V. Induction of p-coumaroyldopamine and feruloyldopamine, two novel metabolites, in tomato by the bacterial pathogen *Pseudomonas syringae*. Molecular plant-microbe interactions. 2007;20(20):1439-1448.
88. Zhao J. Flavonoid transport mechanisms: how to go, and with whom. Trends in Plant Science. 2015;20:576–585.
89. Zhao J, Huhman D, Shadle G, He XZ, Sumner LW, Tang Y, Dixon RA. MATE2 mediates vacuolar sequestration of flavonoid glycosides and glycoside malonates in *Medicago truncatula*. Plant Cell. 2011;23:1536–1555.
90. Zhao J, Dixon RA. MATE transporters facilitate vacuolar uptake of epicatechin 3-*O*-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and *Arabidopsis*. Plant Cell. 2009;21:2323–2340.

Additional Files

Additional file S1: File S1a-h. Expression levels of *BAHD* genes in different soybean tissues, during nodule developing, Al^{3+} and low pH conditions, fungal elicitors, in different seed coat colour genotypes, hairy roots overexpressing *MYB3*, *MYB7*, *GmLEC2a* and *GmWRi1b* .

Additional file S1: File S2. List of primesr of soybean BAHD genes used for qRT-PCR validation.

Additional file S1: File S3. Physical parameters of soybean proteins coded by BAHD genes.

Additional file S1: File S4. Description of motifs in BAHD genes sequences in the soybean genome.

Additional file S1: File S5. *Cis*-acting promoter region analysis of soybean BAHD genes.

Additional file S2: Table S1. Duplicated BAHD genes in soybean and the estimated dates of their duplication event.

Additional file S2: Table S2. Maximum likelihood analysis of BAHD proteins for codon by codon positive selection.

Additional file S2: Table S3. Mixed-effect model evolution (MEME)-based episodic diversifying selection BAHD genes.

Additional file S2: Table S4. Fast unconstrained Bayesian approximation inferring pervasive diversifying selection of BAHD genes.

Additional file S2: Table S5. Codon model selection-based on Modified Bayesian Information Criterion (mBIC) of BAHD genes.

Additional files S3: Figure S1. Cumulative behaviour of synonymous, non-synonymous and ambiguous codon mutation.

Additional files S3: Figure S2. Expression of selected BAHD in transgenic soybean hairy roots overexpressing *GmMYB3* and *GmMYB7*.

Figures

Fig. 1

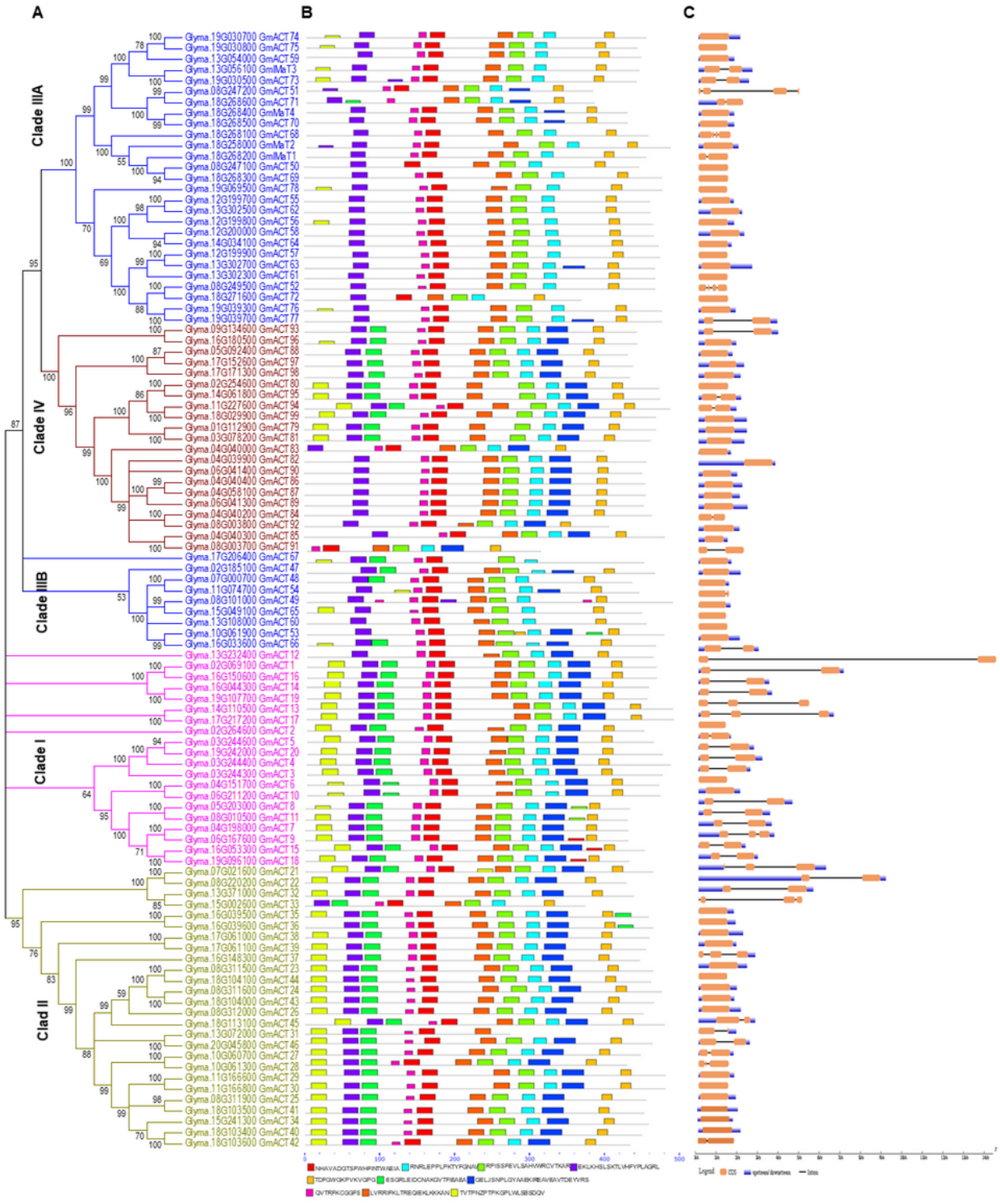

Figure 1

Phylogenetic analyses, conserved motif and gene structure analysis of soybean BAHD family genes. A) Phylogenetic tree and nomination of all soybean BAHD family genes (103 in total). The genes are listed according to their fall in the clade and position on the chromosome and different clades are highlighted in different colors: Clade I in fuchsia, Clade II in olive, Clade III in blue and Clade IV in maroon. B) The motifs of soybean BAHD proteins are shown as colored boxes. The GmACT proteins are listed according

to the order of clades I-IV from the phylogenetic tree. C) The gene structures of 103 GmACTs were plotted using light orange boxes representing exons (coding DNA sequence), black lines representing introns and blue boxes indicating UTR sequences. The scale in the bottom is in the unit of kilobase (kb).

Fig. 2.

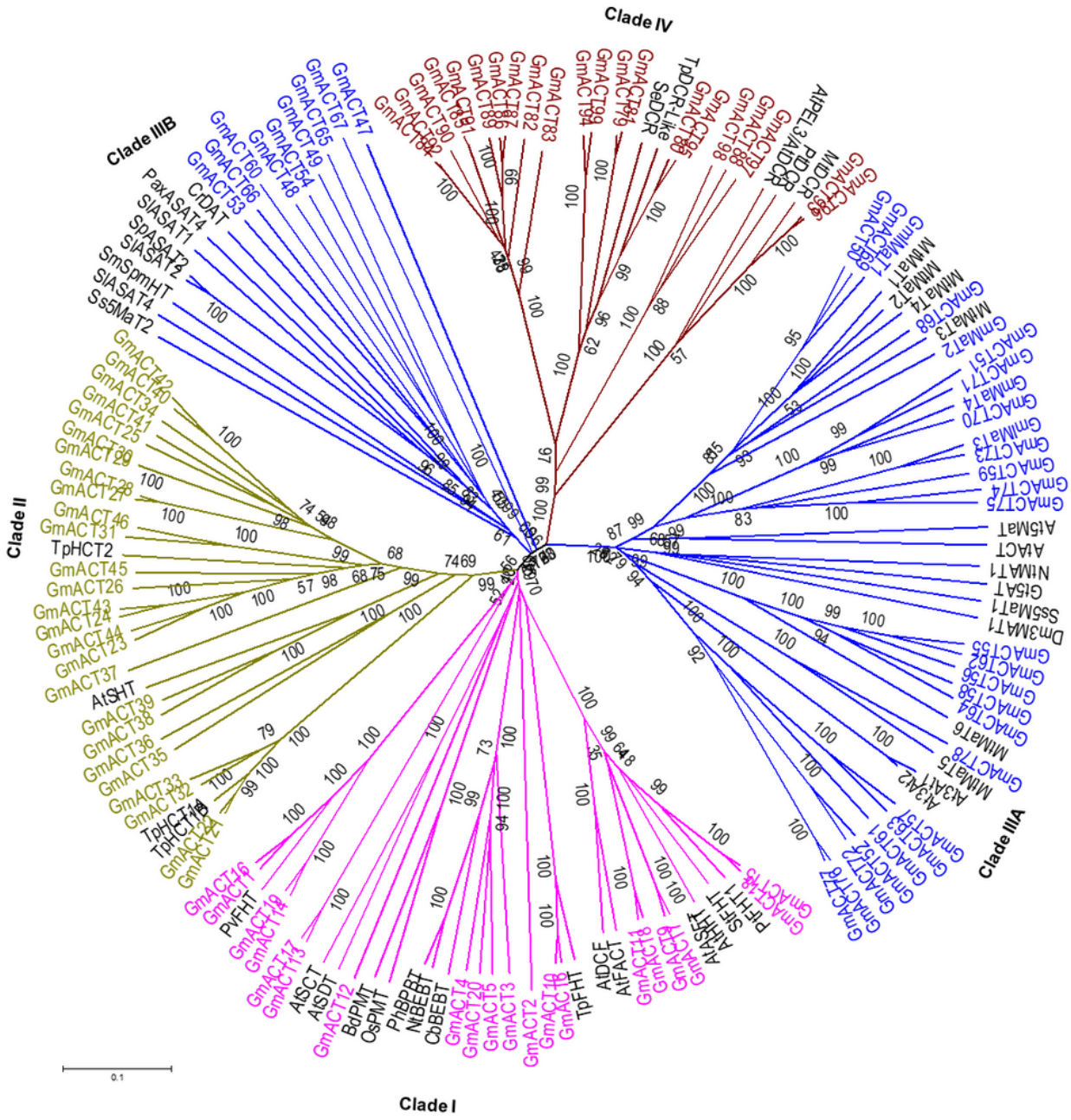

Figure 2

Phylogenetic tree of soybean BAHD family genes with characterised BAHD genes from different plant species. The phylogenetic tree was constructed through MEGA 6.0 using the Maximum Likelihood (ML)

method. Bootstrap values in percentage (1000 replicates) are indicated on the nodes. Different clades are highlighted using different colors (same as Figure 1A).

Fig. 3.

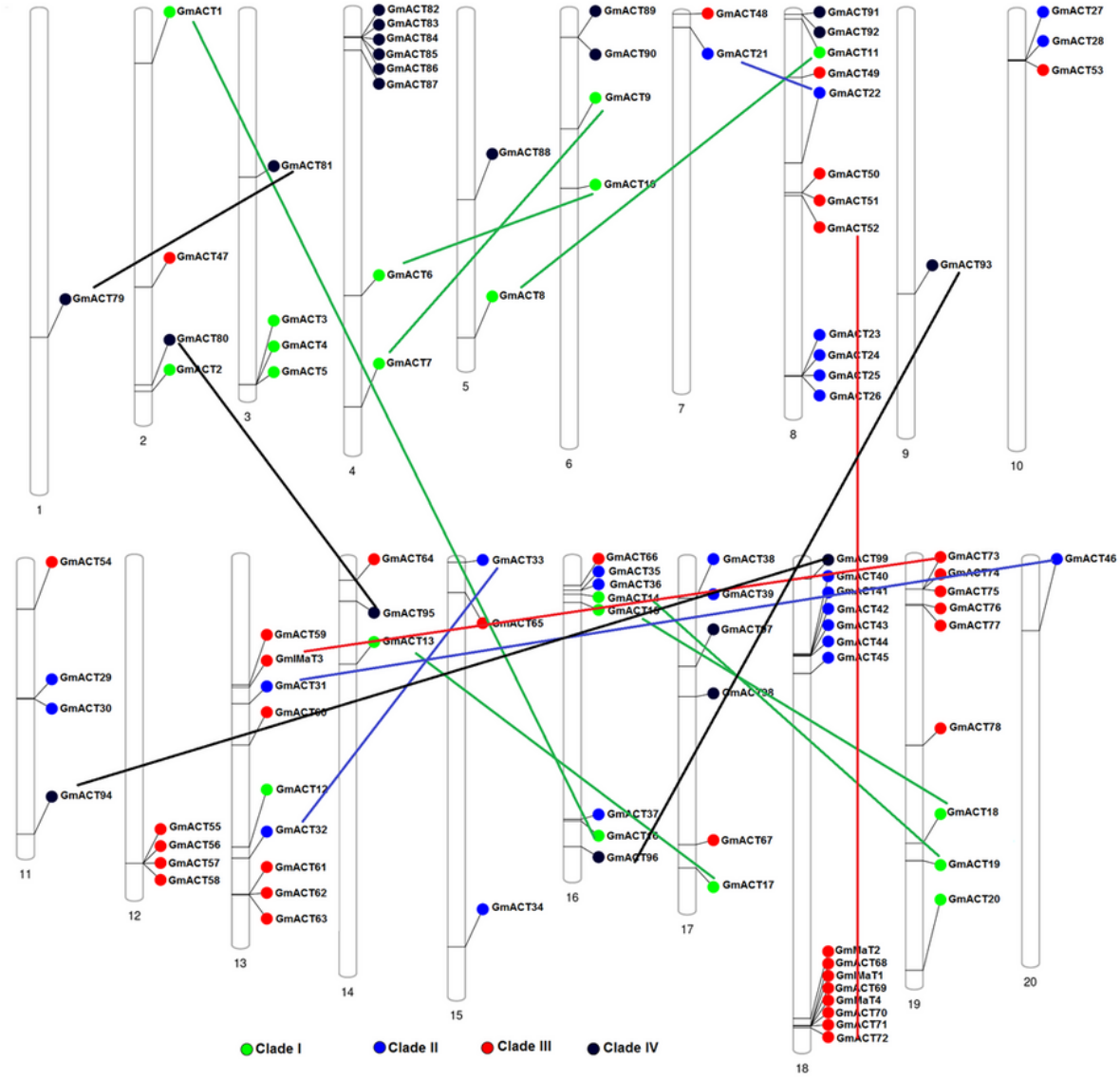

Figure 3

Chromosomal locations and duplications of soybean BAHD genes. The chromosome number is indicated below each bar. The chromosome size is indicated by its relative length using the information from Phytozome and SoyBase. The different coloured circles represent the genes belong to specific clade on

each chromosome (green from clade I, blue from clade II, red from clade III and black from clade IV). Each pair of segmental duplication is indicated by a respective clade color line.

Fig. 4.

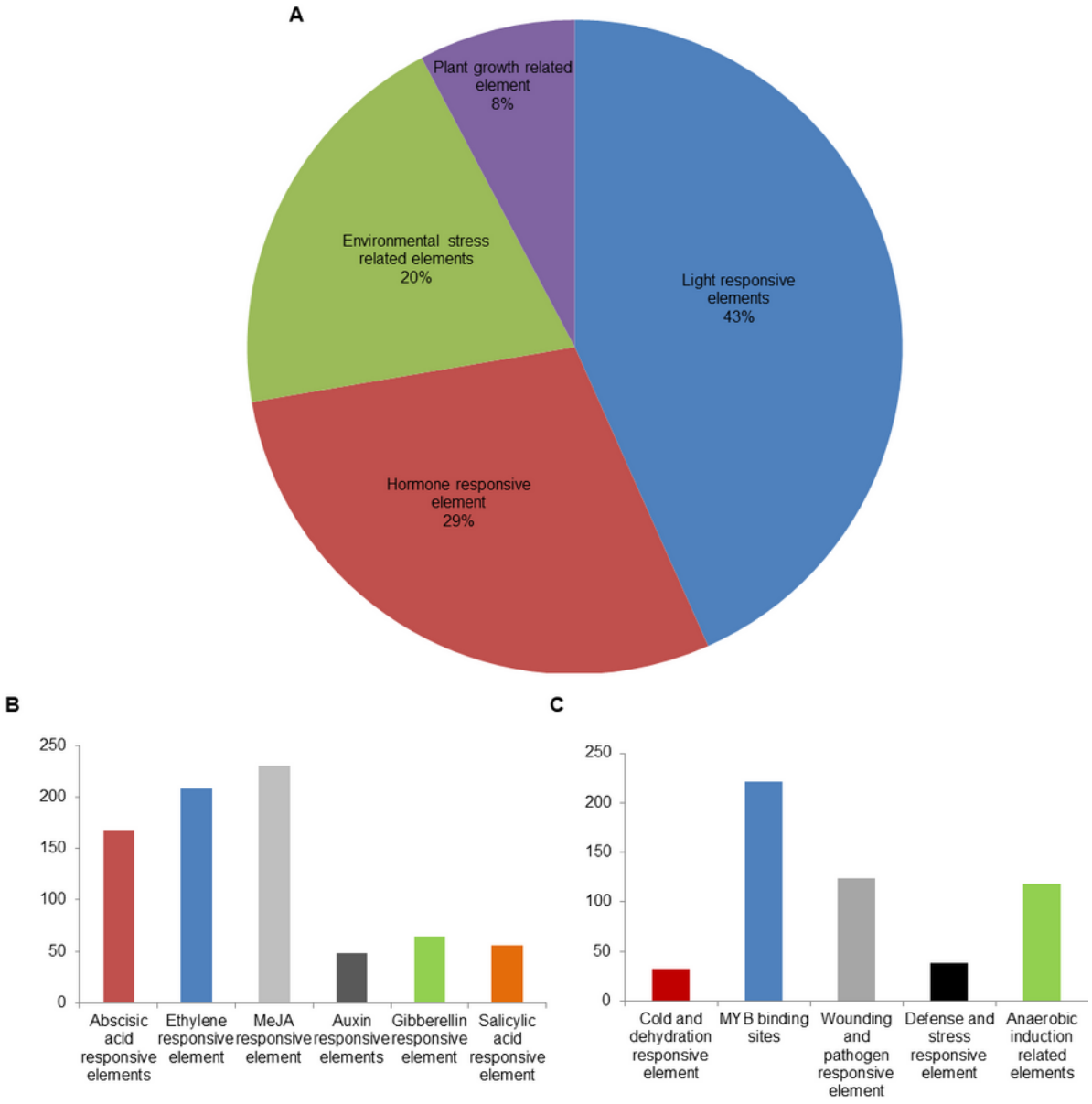

Figure 4

Cis-element analysis in the soybean BAHD family. PlantCare were used to analyze the region 1500 bp upstream of each of the GmACT gene. A) The percentage of light responsive elements, hormone responsive elements, environmental stress related elements, and plant growth responsive elements in all

BAHD family members. B) Different hormone (ABA, ethylene, MeJA, auxin, gibberelin, salicylic acid) responsive elements in 103 BAHD genes cis-element regions. C) Different environmental stress (heat, cold and dehydration, drought, defence, anaerobic, wound and pathogen) related elements in 103 BAHD genes cis-element regions.

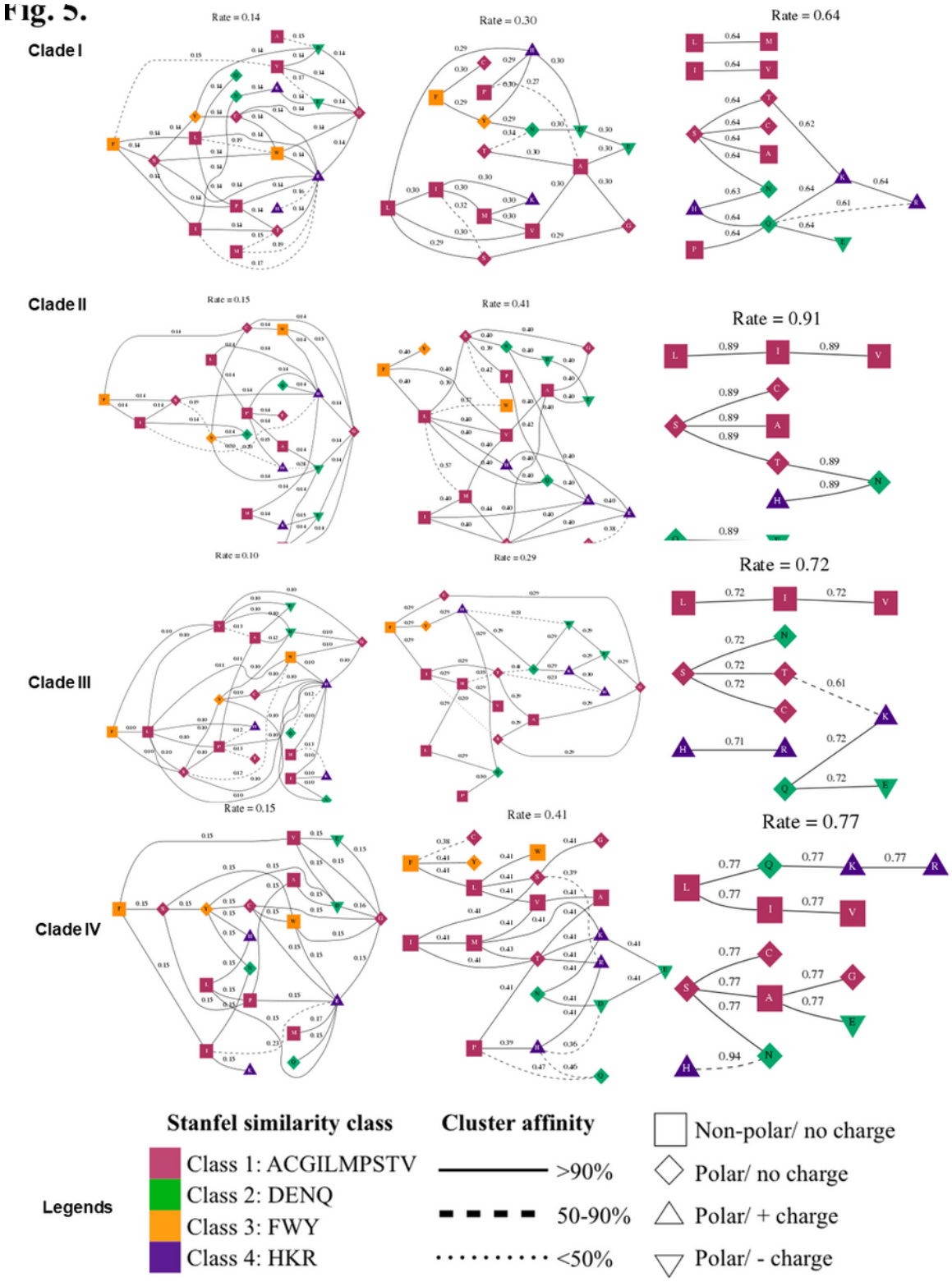

Figure 5

Evolutionary rate cluster in structured genetic algorithm models (GA) inferred from GmACT gene alignment of different clads. Each cluster labelled with maximum-likelihood estimate of its rate inferred under genetic algorithm. The nodes (residues) are annotated by their biochemical properties and Stanfel class, and the rates (edges) are labelled with model averaged rate estimates.

Fig. 6.

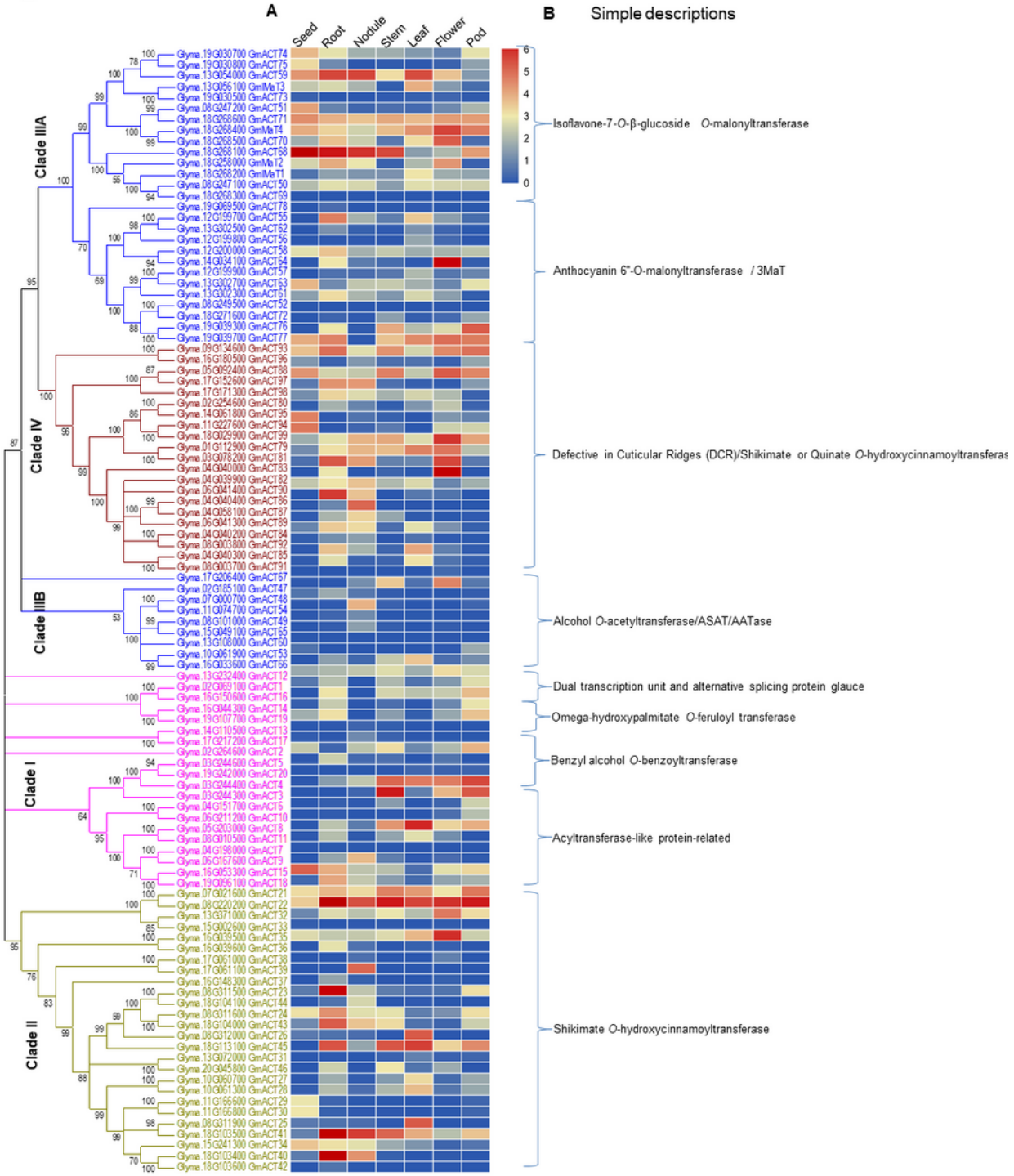

Figure 6

Expression patterns of BAHD family genes in different soybean tissues The expression level of different BAHD family genes in different soybean tissues were retrieved from Phytozome database and represented by constructing the heat map using TBTools program.

Fig. 7.

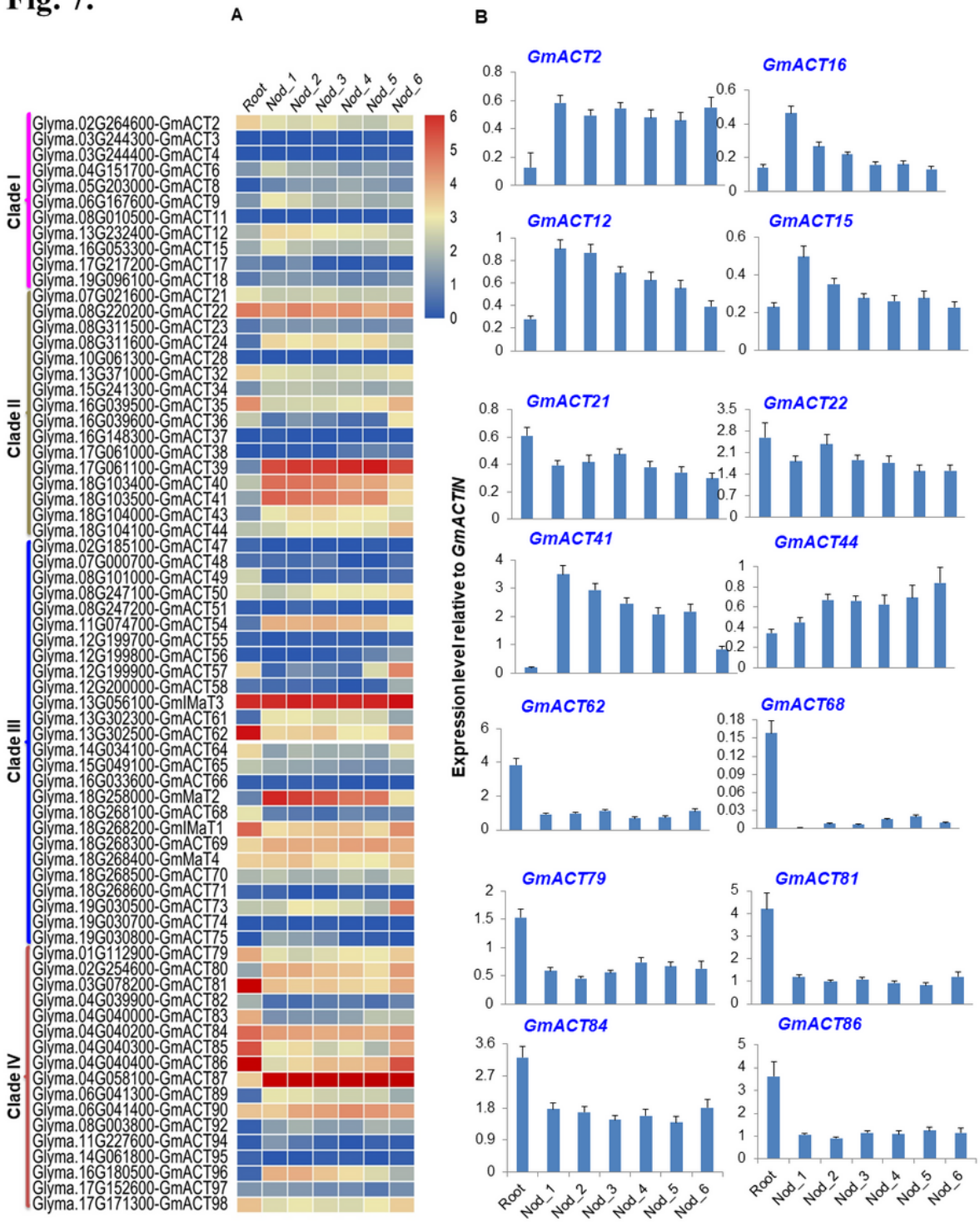

Figure 7

Expression pattern of BAHD family genes in different soybean nodule developing stages The expression level of different BAHD family genes in different soybean nodule developing stages were retrieved from

generated Solexa sequencing libraries, analyzed and represented by constructing the heat map using TBTools program. qRT-PCR analyses of selected GmACT genes in different nodule developing stages were done. N-1, Nodule developing stage-1; N-2, Nodule developing stage-2; N-3, Nodule developing stage-3; N-4, Nodule developing stage-4; N-5, Nodule developing stage-5; N-6, Nodule developing stage-6.

Fig. 8.

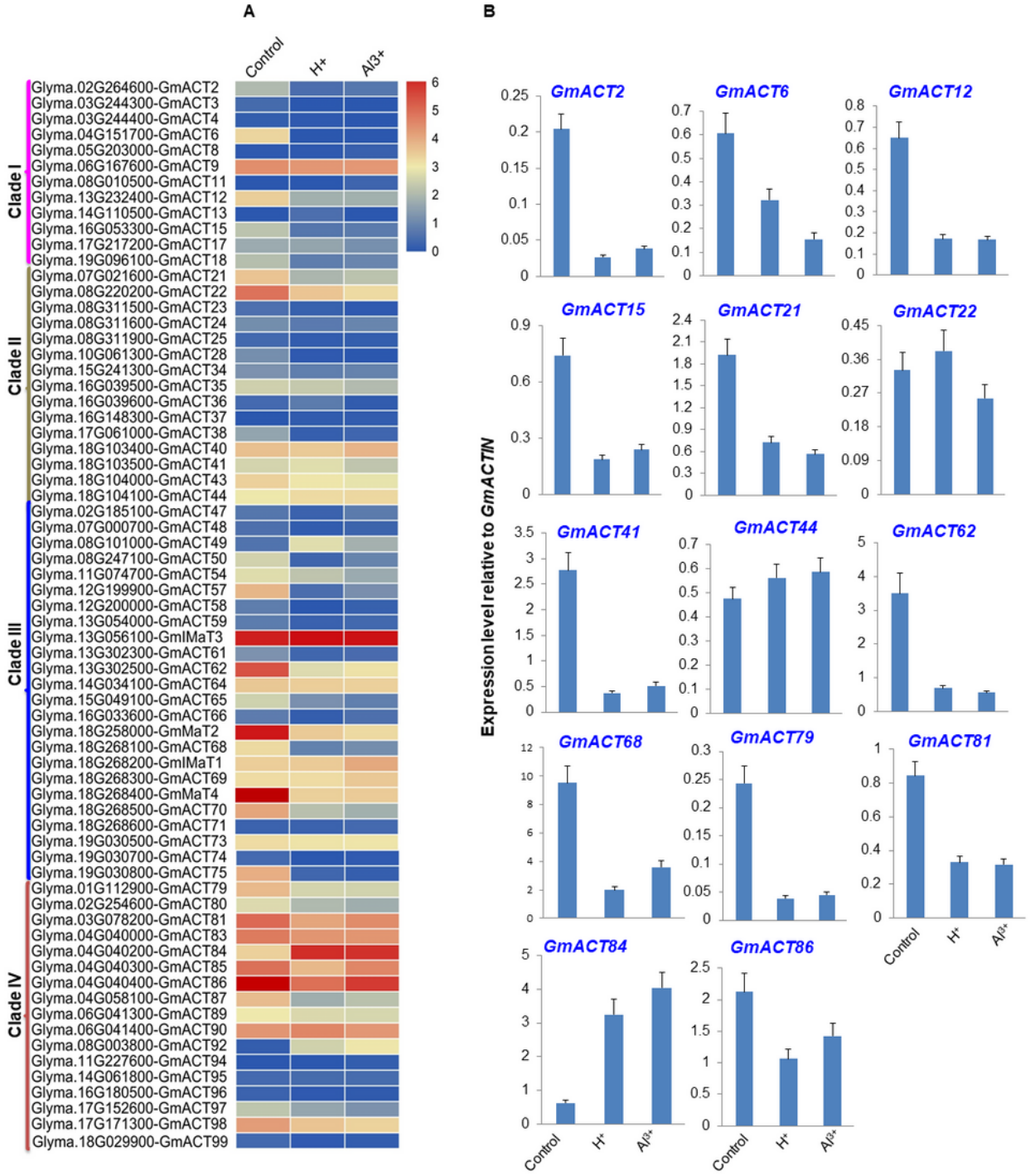

Figure 8

Expression pattern of BAHD family genes in response to low pH and Al3+ stress The expression level of different BAHD family genes in in response to low pH and Al3+ stress were retrieved from generated Illumina sequencing libraries, analyzed and represented by constructing the heat map using TBTools program. Soybean seeds were germinated in vermiculite in a light chamber at 25 ± 2 ºC for about 2 weeks. For acidic condition (pH 4.0) treatment and 50 mM Al3+ stress (under pH 4.0), hydroponically cultivated seedlings were transferred to these media for 10 days before harvesting roots for analysis. The roots were collected for gene expression analysis.

Fig. 9.

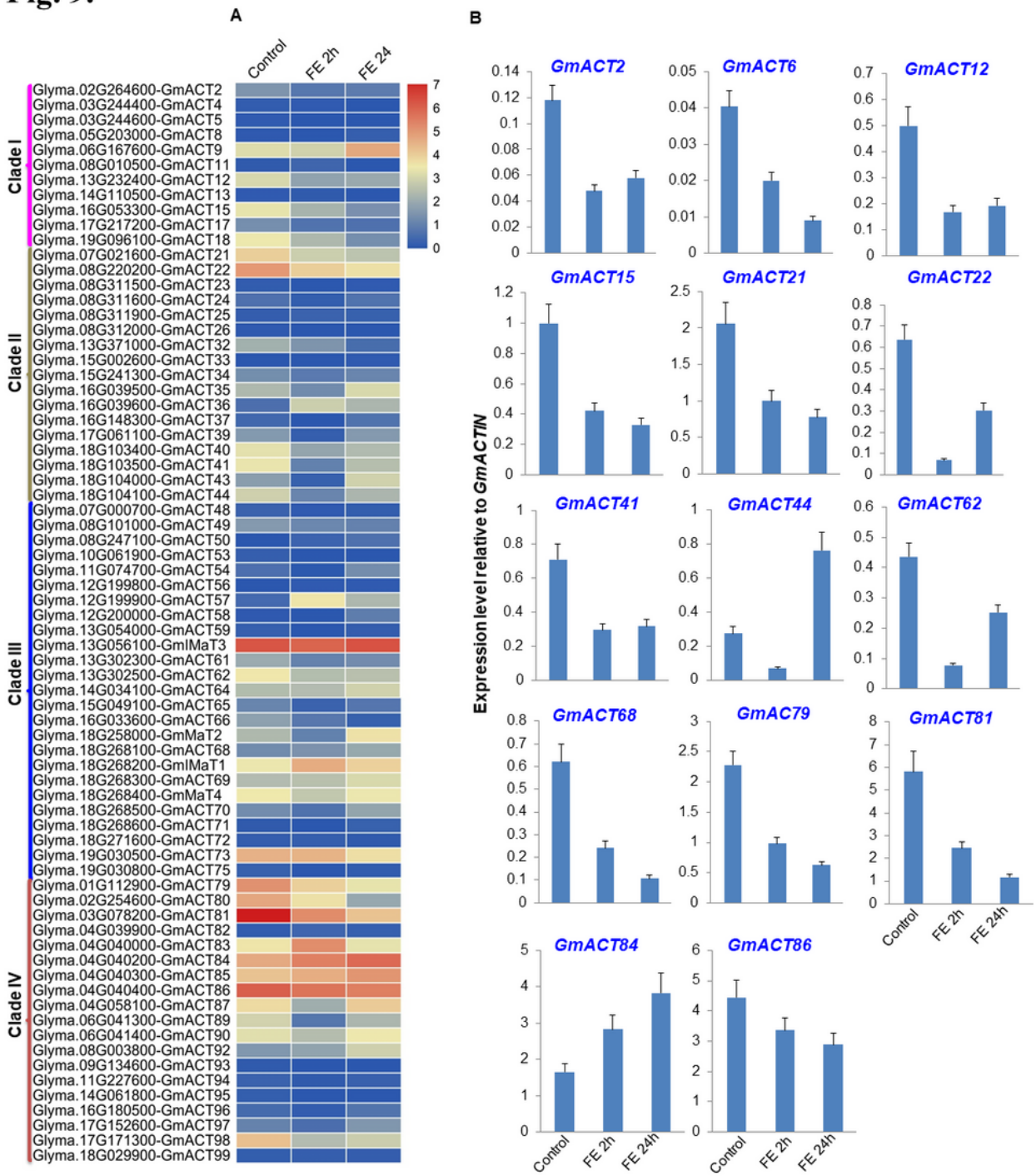

Figure 9

Expression pattern of BAHD family genes in response to fungal elicitor The expression level of different BAHD family genes in in response to fungal elicitor were retrieved from generated Solexa sequencing libraries, analyzed and represented by constructing the heat map using TBTools program. The Fusarium spp. oligosaccharide elicitor was prepared. The phenol/sulphuric acid method was used to determine the concentration of the elicitor. Seedlings were transferred into a hydroponic culture system. The soybean seedlings were grown in an incubator at 28°C with a 16/8 h photoperiod for 24 h. For elicitor treatment, these wild-type plants were transferred into hydroponic medium and grown under controlled condition for 14 days just same as normal soybean plants. Elicitors (30 mg/ml fungal elicitor or 50 μM MeJA) or inhibitors were added to solution for treatments. Roots were collected at 2, and 24h after treatment for gene expression analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile3.ppt
- Additionalfile2.doc
- Additionalfile1.xls